The sensitivity and specificity of patient-specific QC at the Wellington Blood and Cancer Centre

A thesis submitted in partial fulfilment of the requirements for the

Degree of Master of Science in Medical Physics

by

Benjamin Ewen William Scarlet

Department of Physics and Astronomy

University of Canterbury

2017

Abstract

Purpose
Patient-specific quality control (QC) plays an important role in assuring the safety of treatment planning and delivery for complex treatment techniques such as volumetric modulated arc therapy (VMAT). Ideally, patient-specific QC should be able to detect clinically relevant errors in treatment plans (sensitivity), pass treatment plans that do not contain errors (specificity), and resolve different error modes. Previous studies in literature have reported that patient-specific QC methods have a lower sensitivity than specificity. Therefore, the aim of this study was to quantify the sensitivity and specificity of the patient-specific QC methods currently available at the Wellington Blood and Cancer Centre. A secondary aim was to determine if any QC method could resolve different error modes.

Methods
Intentional errors simulating incorrect linac monitor unit delivery (MU), multi-leaf collimator (MLC) positioning, dosimetric leaf gap (DLG), focal spot size (FSS) and output variation with gantry angle were introduced to five Head & Neck VMAT plans. Criteria were defined to determine whether each introduced error caused a clinically relevant dose deviation to the patient treatment. Error-free plans and introduced error plans were measured using a time resolved point dose method (trPD), an EBT3 film method (Ashland Inc.), and using an ArcCheck helical array (Sun Nuclear corp.). Sensitivity of each QC test (true positive rate) and intrinsic measurement system sensitivity (change in output over change in input) were calculated, as well as the specificity of each QC test (true negative rate). In addition, receiver operator characteristic (ROC) curves were created for each QC method, and the efficiency of each QC method was determined by calculating the area under the ROC curve (AUC). In addition, the ability to resolve different error modes was investigated for the trPD method.

Results
A total of 89 plans were created (5 error-free, 84 containing introduced errors). Of the 84 introduced errors, 52 caused clinically relevant dose deviations. Using the clinically applied QC acceptance criteria (±2% dose difference for trPD, >85% of points passing a {2%;2mm} γ-criterion for film and ArcCheck), all three QC methods were found to have high specificity but the sensitivity was comparatively low (**Table 1**, first row).By varying the acceptance criterion for trPD to ±1.9%, and modifying the beam model and acceptance criterion to 87% for film, the sensitivity of these methods could be improved at the expense of a slight decrease in the specificity. However, the observed improvements in efficiency were within the estimated uncertainty range. For the ArcCheck system, none of the investigated configurations yielded an acceptable sensitivity and specificity (**Table 1**, second row, optimal practise involved varying the passing criterion to 92%). Investigation of the ArcCheck intrinsic sensitivity indicated that these poor results were caused by a systematic offset between the measured and TPS calculated dose. Analysis of the trPD results showed that different error modes could potentially be resolved by using a per region analysis, where the detector distance to the MLC field edge defined the different regions.

Table 1: *Sensitivity (true positive rate), specificity (true negative rate) and efficiency (in terms of the AUC) for the three QC methods using current clinical practise and using QC passing criteria based on ROC analysis.*

	Point Dose			Film			ArcCheck		
	Sensitivity	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**
Current practise	65%	91%	0.79	61%	83%	0.77	47%	80%	0.74
Optimal practise	67%	89%	0.79	80%	71%	0.80	76%	64%	0.74

Conclusion
The current patient-specific QC methods at the WBCC displayed a low sensitivity for clinically relevant errors but a high specificity, similar as reported by other published studies. This study showed the importance of quantifying the characteristics of patient-specific QC methods in more detail prior to clinical application.

Acknowledgements

Firstly I would like to thank both my clinical supervisors Drs. Andrew Williams and Rob Louwe for their on-going support, insight and ideas that were a huge contribution to this work. I would also like to thank them for persevering through many confusing meetings as I struggled to explain my ideas and progress – a huge thanks to you both! I would also like to thank my academic supervisor Dr. Steven Marsh for his support, encouragement and input from afar at the university. Thanks Steve!

I would like to acknowledge the large amount of support I receive from Ms. Lynne Greig and the Wellington Blood and Cancer Centre for providing the time and resources necessary to conduct this project. I would also like to give a big thank you to all the other medical physicists at the WBCC who were always happy to help out if I had any questions or provide their own thoughts and advice.

A huge thank you my family, friends, and flatmates for your never wavering encouragement and putting up with my constant talk about this project even when away from it. All your support cannot be understated and I am really grateful for all your help! Finally, I would like to give a special thank you to Amanda, for always being there for me to talk to and helping me through the most stressful parts of this project. I couldn't have completed this work without you.

Table of Contents

Glossary

3D-CRT	*3-Dimensional conformal radiotherapy*
AAA	*Analytical anisotropic algorithm*
AUC	*Area under curve*
BS	*Brainstem*
CI	*Confidence interval*
CP	*Control point*
CTV	*Clinical Target Volume*
DICOM RT	*Digital imagining and communications in medicine file format for radiation therapy*
DLG	*Dosimetric leaf gap*
DTA	*Distance to agreement*
DTFE	*Distance to field edge*
DVH	*Dose-volume histogram*
DVO	*Dose-volume objective*
FN	*False negative*
FP	*False positive*
FSS	*Focal spot size*
GTV	*Gross tumour volume*
H&N	*Head and neck*
HU	*Hounsfield units*
ICRU	*International commission on radiation units and measurements*
IMRT	*Intensity modulated radiation therapy*
Linac	*Linear accelerator*
MLC	*Multi-leaf collimator*
MU	*Monitor units*
NTCP	*Normal tissue complication probability*
OAR	*Organ at risk*
OD	*Optical density*
PPTV	*Planning target volume (for VMAT optimisation)*
PRO	*Progressive resolution optimiser*
PRV	*Planning risk volume*
PTV	*Planning target volume*
QA	*Quality assurance*
QC	*Quality control*
ROC	*Receiver operator characteristic*
S_1	*Sensitivity metric 1*
S_2	*Sensitivity metric 2*
S_3	*Sensitivity metric 3*
SC	*Spinal cord*
Sp_1	*Specificity metric 1*
Sp_2	*Specificity metric 2*
TCP	*Tumour control probability*
TN	*True negative*
TP	*True positive*
TPS	*Treatment planning system*
trPD	*Time resolved point dose*
(E)TSS	*(Effective) target spot size*
VMAT	*Volumetric modulated arc therapy*
WBCC	*Wellington Blood and Cancer Centre*

1. Introduction

Cancer is one of the leading causes of mortality worldwide with approximately 14 million new cases diagnosed and 8.2 million cancer related deaths occurring in 2012 [1]. In New Zealand, cancer is the leading form of mortality, accounting for 29.4% of all deaths in 2011. In the same year there were 21,050 new cases of cancer registered and 8,891 deaths due to cancer [2]. The three most common methods to treat cancer are surgery, chemotherapy and radiotherapy. These three treatment modalities can be used individually or combined depending on the location of the cancer, the tumour staging and the intent of the treatment i.e. is it curative or palliative. Approximately 50% of all patients being treated for cancer worldwide will receive radiotherapy as part of their treatment, as either their primary form of treatment, or in conjunction with other treatment modalities [3]. The most common form of administering radiotherapy is through external beam radiation therapy (EBRT) where a high energy electron or photon beam (or multiple beams) produced by a linear accelerator is used to treat the patient.

For radiotherapy to be an effective treatment for cancer, it is required that the targeted volume (or volumes) is correctly irradiated to a dose level (given in joules.kg^{-1} or Gray) that has been prescribed by a radiation oncologist (RO). The main aim of radiotherapy is to stop the tumour growing, and then if possible, completely eradicate the tumour from the patient by killing all the tumour cells. The probability of achieving this is known as the tumour control probability (TCP) [4]. However, radiation is not selective in which cells it targets, and therefore the probability of radiation damage to normal tissue must also be taken into account through the normal tissue complication probability (NTCP) [5]. Therefore, the NTCP of the surrounding organs at risk (OARs) must be taken account of and balanced against the TCP. The entire premise of radiotherapy is achieving a high TCP while maintaining a low NTCP and this is known as the therapeutic index (see **Figure 1.1**) [6]. In order to access the

therapeutic index and hence the quality of a radiation treatment plan, it is necessary to look at how the dose is distributed over the target compared to the OARs on a computer simulated 3D image of the patient. One method to obtain a clinically acceptable therapeutic index is through the use of splitting the course of a treatment up into a large number of sessions called fractions. This allows late responding tissues (such as many OARs) to repair themselves following exposure to radiation, while still causing irreparable damage to early responding tissues (most tumours) [7]. This practise is common for most radiotherapy treatments. The therapeutic index can be further improved by optimising the radiation beam arrangements to produce a treatment plan in which more dose is absorbed by the target while less dose is absorbed by surrounding OARs i.e. producing a more conformal dose distribution [8].

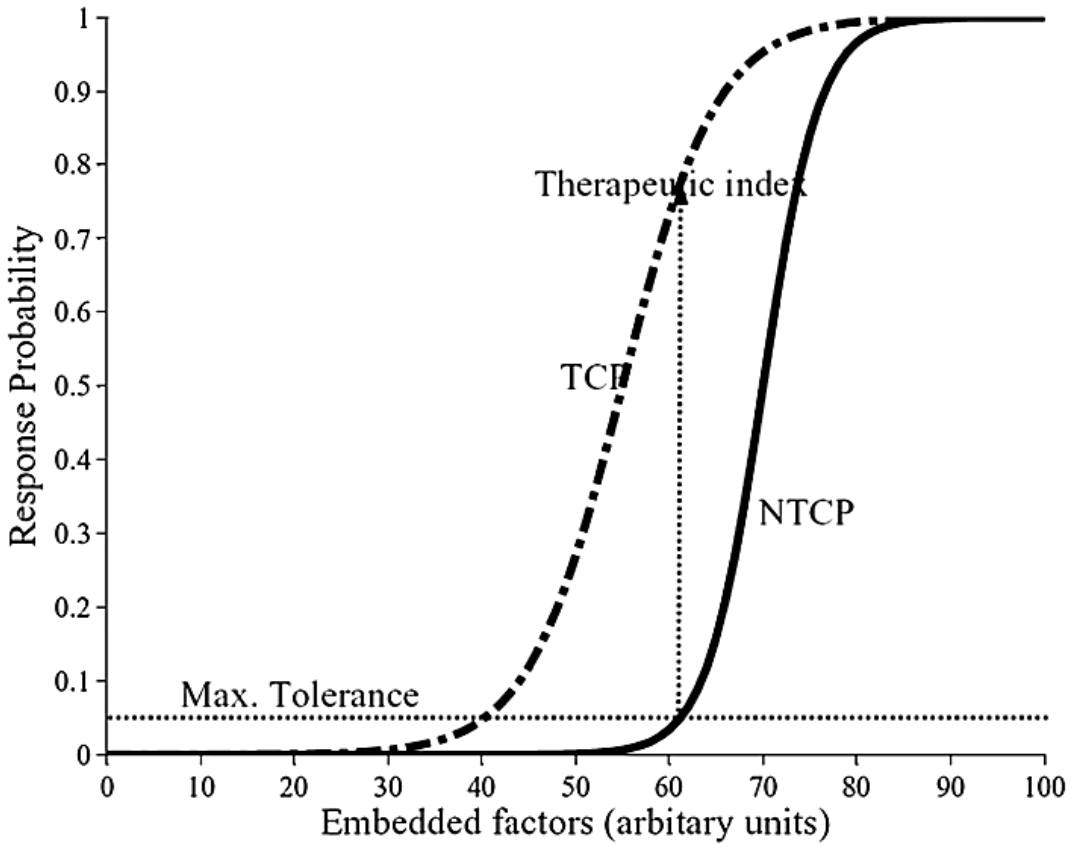

Figure 1.1: *TCP and NTCP curves indicating the therapeutic index for the maximum tolerance of a particular normal tissue. Figure reproduced from El Naqa et al.* [9].

Radiotherapy technology and treatment methods are continually evolving to improve the outcomes of patients undergoing cancer treatment. These new treatment methods increase the likelihood of

controlling disease while reducing undesirable side effects through delivering highly conformal dose distributions. However, in order to achieve this goal, a high degree of control and accuracy in the delivery of radiotherapy plans is necessary in order to ensure that these highly conformal plans can be delivered to the patient accurately and without errors occurring during the course of their treatment. This requires not only a high level of delivery accuracy, but also treatment plan calculation accuracy, computer technology to warrant correct transfer of plans from the treatment planning system to the treatment machine accurately, and the development of meaningful quality assurance tests to ensure every step of the process is conducted accurately.

1.1. 3D Conformal Radiation Therapy

The majority of radiotherapy treatments since the early 2000s have been delivered using three-dimensional conformal radiotherapy (3D-CRT). The basic idea of this technique is to shape the beam aperture of a photon beam generated by a linear accelerator (linac) to closely conform to the shape of the target volume. This allows for the ability to deliver the prescribed radiation dose to the target volume, while reducing the dose delivered to surrounding tissue, thus minimising the adverse effects of treatments [10]. Historically, 3D-CRT was delivered using lead shielding blocks to shape the beam aperture. These had to be designed and made individually for each beam of each patient treatment. In practice, this limited the number of beams that could be delivered, led to long treatment times due to the need to change lead blocks for each beam, and posed a risk of incorrect dose delivery if the wrong blocks were inserted. In the last decade of the twentieth century the use of lead blocks began to be replaced by multi-leaf collimators (MLCs) for delivering 3D-CRT. This device enables continuous modification of the shape of treatment beams' collimation and was implemented on a large scale at the end of the previous century [11-12]. An MLC consists of two banks of abutting tungsten leaves (see **Figure 1.2**) with each leaf being capable of independent movement. This allowed the photon beam aperture to be shaped almost instantaneously to that of the target volume and also to shield healthy tissue and organs at risk.

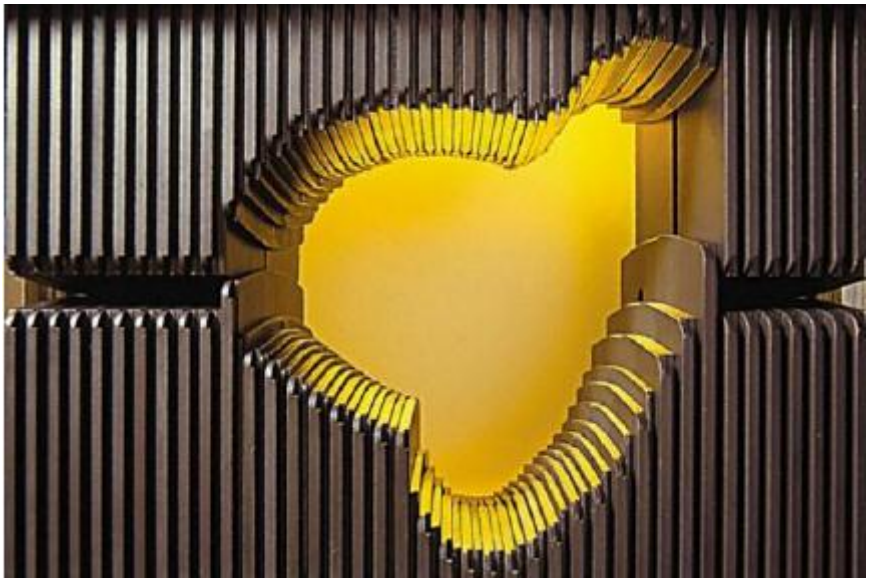

Figure 1.2: *Image of an MLC with 60 pairs of leaves. Reproduced from Oliveira et al*. [13].

3D-CRT treatment planning is carried out using the following process:

1) Anatomical data from a multiple slice computed tomography (CT) scan is used by the RO to identify and contour the target structures and OARs.
2) Planning staff create the treatment plan using a computer based treatment planning system (TPS). This is usually a manual process whereby field gantry angles are selected, the field aperture defined, and the field weight and wedge angle (if used) are set. The plan will usually involve multiple beams delivered using different gantry angles.
3) The dose to the patient is calculated by the TPS. It determines the dose to the 3D patient volume using appropriate dose calculation algorithms and a properly configured beam model (see section 1.3).

Steps 2) and 3) are iterative and are repeated until the optimal dose distribution is obtained. The planning process requires planning staff with a high level of expertise. However, 3D-CRT does have its limitations. 3D-CRT was designed for use with lead blocks and even though it is now conducted using MLCs, the MLC positions remain static, and the geometry is generally limited to a low number of gantry angles (although some treatments do utilise a high number of beams at various gantry

angles). Therefore only a limited level of conformity can be achieved in many cases, particularly if an OAR is in close proximity to a target structure. Furthermore, since most 3D-CRT plans require multiple beams to achieve an acceptable level of conformation, the overall treatment time can be long, as the patient needs to be set up for treatment and further time is required to rotate the linac gantry between beams.

1.2. Intensity Modulated Radiation Therapy

Intensity modulated radiation therapy (IMRT) is a form of radiation treatment which is capable of providing a higher level of dose conformity than 3D-CRT. While IMRT could be delivered using physical compensators and wedges, the invention of the MLC was essential for the development and widespread adoption of IMRT [8, 14]. Similarly to 3D-CRT, the beam aperture is conformed to the target volume using an MLC. However in 3D-CRT, the photon fluence is homogeneous across the beam aperture, whereas for IMRT, the photon fluence is heterogeneous across the beam aperture. The photon fluence is able to be varied using the highly accurate positioning of computer controlled MLCs. This allows an even higher level of dose conformation to the target volume and/or normal tissue sparing compared with 3D-CRT. IMRT can be delivered using two different techniques. These are:

1) Static MLC IMRT (SMLC-IMRT, also known as step and shoot IMRT), in which each radiation beam at a particular static gantry angle is made up of a number of individual segments. Each segment in turn consists of a static MLC-defined pattern with a uniform fluence. Upon delivery of each segment, the radiation beam is turned off and the MLC leaves move to form the aperture defined in the following segment, thus building up a beam of heterogeneous fluence [8, 15].
2) Dynamic MLC IMRT (DMLC-IMRT, also known as sliding window IMRT), in which the gantry remains static during delivery of each beam (as for SMLC-IMRT), however,

the MLC leaves are allowed to move, with the MLC leaves producing the desired non-uniform fluence though constant motion during the delivery of each beam [8, 15].

For both SMLC-IMRT and DMLC-IMRT, the combination of multiple beams, which individually give a non-uniform dose to a target volume, combine to provide a highly conformal uniform dose distribution [16].

Both of the above methods are capable of delivering much more conformal treatments to the patient than 3D-CRT. Furthermore, IMRT generally uses an inverse planning process (see section 1.3.2). This involves the planner setting optimisation constraints on the treatment plan (the dose to the target structures and the dose constraints on the OARs), and then allows a computer to optimise the MLC movement to achieve the desired outcomes which helps achieve a more conformal dose distribution.

1.3. Treatment Planning

1.3.1. Beam Modelling

Traditionally, patient dose distributions were calculated using measurement based corrections i.e. making a dose measurement in a phantom and correcting it for the different conditions between the water phantom and the patient. Whereas modern 3D-CRT and IMRT utilise model-based algorithms to conduct dose calculations. Model based algorithms determine the dose distribution from first principles by providing a model of the photon beam source and predicting how the generated photons interact. However, model-based algorithms still require the collection of measurement data to set up some model parameters and to verify the beam model [17]. This involves collecting measurement data from a linac radiation beam, inputting this data into the TPS, and configuring this data such that it can be used to accurately calculate dose distributions in patients. This process is conducted during the initial commissioning of a linac or TPS. Data to be collected includes radiation beam profiles (for both jaw defined and MLC defined fields) for multiple field sizes and depths, percentage depth doses

or PDDs (a measure of the radiation dose deposited as a function of depth in a medium) for various field sizes, scatter factors (a measure of the amount of dose contributed from scattered radiation), and accessory factors (for example wedge factors which account for the difference in dose with a wedged field compared to a non-wedged field). Once this data has been input into the TPS and configured, testing should be conducted in the form of calculating treatment plans using the TPS and then measuring the dose delivered using this treatment plan and comparing to the dose calculated. These verification tests should be carried out using simple plans, as well as plans that are of similar complexity to those that will be used for clinical treatments. Beam modelling plays a very important role in radiotherapy to ensure the correct plan is generated to deliver the prescribed dose to the patient.

Although the complexity of radiation treatments and beam modelling has increased from 3D-CRT to IMRT, it is difficult to commission a TPS for all possible combinations of parameters used in IMRT deliveries. Therefore the verification of beam modelling for IMRT deliveries often becomes a part of patient-specific QA (see section 1.7.2).

1.3.2. Forward Planning Vs. Inverse Planning

Forward planning is the process of a planner specifying the arrangement of treatment beams (including positioning of MLC leaves and use of enhanced dynamic wedges and other beam modifiers etc.). The planner will then calculate the resultant dose distribution and then check whether the dose distribution adequately covers the target volume and spares all the relevant OARs, iteratively modifying the beam arrangement and beam modifiers until a satisfactory dose distribution is obtained. In contrast, inverse planning is the process of specifying the desired dose to the target volume and the dose constraints of the OARs. These are known as the dose-volume objectives (DVOs). Once these have been specified, the TPS will iteratively optimize the treatment parameters until the specified dose-volume objectives are met, while keeping the treatment fields within a predefined range of

operation that is physically achievable [18]. In theory, it would be possible for a treatment planner to use forward planning to result in the same treatment beams as the optimization process. However, in reality this would be a very time consuming process as the resultant beam apertures for each particular segment are generally not intuitive.

1.4. Volumetric Modulated Arc Therapy

An IMRT treatment consists of a number of beams at static gantry angles which each consist of a number of segments. This then led to the development of Intensity Modulated Arc Therapy (IMAT), in which the MLC leaves are able to move as in DMLC-IMRT, but the gantry is also able to rotate during delivery, resulting in the treatment being delivered in an arc around the patient. The treatment would then consist of a number of partial arcs (each delivered using a constant dose rate) around the patient, with each arc delivering radiation to a different intensity level [8, 19]. This concept was extended further to develop volumetric modulated arc therapy (VMAT). VMAT consists of a single arc broken up into a number of segments known as control points, where each control point has its own MLC aperture *and* dose rate. This allows the delivery of multiple intensity levels in a single arc, unlike in IMAT (in which the intensity level is constant throughout the arc). The individual MLC leaves move in synchronization with the gantry rotation and the dose rate fluctuations to reach the specified MLC aperture and dose rate at each control point. VMAT can be considered as an extension of IMRT with more degrees of freedom. However, instead of delivering multiple beams at static gantry angles, a VMAT treatment can be delivered in a single arc of up to 360° [20]. Furthermore, as the treatment can be delivered in a single arc or multiple arcs, the overall treatment delivery time is often significantly shorter than for either IMRT or 3D-CRT [21]. Similarly to IMRT, VMAT plans are created using an inverse treatment planning technique. As more degrees of freedom are available for VMAT, an intricate interplay and synchronization between various parts of the linac is necessary to achieve accurate reproduction of the planned dose distribution during VMAT delivery. Specifically, it is very important to guarantee a high level of accuracy and synchronisation of the gantry angle, MLC leaf position and dose rate variation to ensure the accuracy of VMAT.

1.5. The Necessity of Quality Control in Radiation Therapy

The internationally accepted limitation of delivery of the clinically prescribed dose is ±5% for any type of radiation treatment based on the recommendation of the International Commission of Radiation Units and Measurements (ICRU) report released in 1999 [22]. Before this report was released, the target in radiotherapy was to achieve the prescribed dose to the target to within ±3.5% [23]. Currently, it is generally accepted that the aim for the accuracy of dose delivery in radiotherapy is ±3% [24]. Radiotherapy is an entire treatment chain, starting from when the patient agrees to undergo treatment, extending through their CT scan, structure contouring, treatment planning, plan checking, and set up and verification prior to treatment on the day. Within any of these links in the radiotherapy chain, errors can arise and propagate. These errors may be either random or systematic (see section 1.8.1). Therefore, an accumulation of small errors may lead to an overall clinically unacceptable dose delivery to the patient. As such, the above ±3% margin can be referred to as an 'uncertainty budget' since it takes into account an entire range of uncertainties that are associated with each step within the radiotherapy treatment chain. It is necessary to reduce the uncertainty within each step to the extent reasonably achievable and ensure the uncertainty associated with each step remains low over the lifetime of the particular hardware and/or software associated with that step. This is achieved through the process of quality management. Furthermore, IMRT and VMAT are much more complex treatments than 3D-CRT, with VMAT in particular containing more degrees of freedom that can potentially lead to treatment errors. Therefore, linacs intended for VMAT deliveries need additional checks to ensure they deliver the dose at the required level of accuracy.

1.6. Quality Management

The quality management of radiotherapy can be broken down into three different subsections; quality management, quality assurance and quality control. Quality management refers to the system that maintains the quality of the specific service, in this case, the delivery of radiotherapy. It is not limited to just the technical aspects of treatment delivery, but encompasses clinical, physical and administrative components [25]. Quality assurance (QA) is the planned and systematic actions

necessary to ensure that a product or service will adequately fulfil the requirements for quality. Quality control (QC) is the process through which the actual quality performance is measured. Therefore, the specific quality control tests that are carried out on the radiotherapy TPS and treatment delivery units only make up one part of overall quality assurance [8].

1.7. Linear Accelerator Quality Management Programme

There are a wide variety of difference QC tests that need to be carried out on a linac regularly in order to guarantee its ability to provide the correct treatment to a patient. Such tests are performed daily, weekly, monthly, and annually depending on how critical they are to the day to day running of the linac and the likelihood that the linac performance will change. The relevant tests that need to be carried out are advised in a number of international standards such as IPEM Report 81 [26] and AAPM TG 40 [27]. These international guidelines provide the basic starting point for an effective quality control programme. They outline a range of different QC tests that should be carried out, as well as their frequency and the recommended tolerance to be applied for each test in order to assure the quality of treatment delivery. However, depending on the complexity of treatments in use at a particular radiotherapy department, additional QC will be necessary. Therefore, linacs capable of VMAT will need to undergo various additional QC checks on a regular basis, as the increased demands on the linac performing to specification requires a more stringent quality assurance program.

1.7.1. Routine Linear Accelerator Quality Assurance

For a linac, basic QC checks will range from daily checks such as of constancy of linac output at each clinical energy and checks of safety systems (such as door interlocks and beam off buttons) to more thorough tests conducted monthly and annually. Although the details of the individual QC tests are outside the scope of this study, the overall concept of the QA program is of utmost importance. Routine QA ensures that a number of vital machine performance characteristics are within acceptable tolerances to warrant the accuracy of treatment deliveries. These tests will check for the accuracy of

various parameters such as gantry angle, collimator angle, machine output, MLC positioning, beam symmetry and flatness and others [26].

1.7.2. Patient-Specific Quality Control

The routine QA as described in section 1.7.1 is vital to the safe delivery of treatments on a linac but only evaluates the accuracy of each individual parameter. However, the very irregular, MLC-shaped treatment fields that are used for VMAT are largely outside the scope of the commissioning process that is applied for configuring new TPS models. Consequently, the accuracy of VMAT deliveries has historically not been included in routine machine QA. In particular, these tests do not look at the effect of synchronisation of the various degrees of freedom available to dynamic treatments on the accuracy of the delivered treatment [28].

Most currently available international guidelines do not provide detail on QC tests for VMAT deliveries (although recently some guidelines have been released [29]). Therefore patient-specific QC is used to verify the accuracy of treatment deliveries and aspects of the TPS for individually generated patient treatment plans. There are three main aspects that are verified through the use of patient-specific QC. These are: 1) that the dose distribution in a QA phantom is accurately calculated by the TPS; 2) that the plan is correctly transferred from the TPS to the linac; and 3) that the linac delivers the intended dose within the departmental tolerance levels [29]. Patient-specific QC is vital when a department is developing a new treatment technique, as it is a complete end to end check of the radiotherapy process for that technique that will ensure the treatment machine can deliver the treatment with the required accuracy. This process is carried out before the patient receives their first treatment and is known as pre-treatment patient-specific QC. However, if QA needs to be conducted prior to every patient treatment, resource constraints will limit the number of treatments that are possible. One option to overcome this is to treat every patient and only conduct pre-treatment patient-specific QC on a subset of those patients treated. Or another option is once there is a reasonable

amount of confidence in the technique, pre-treatment patient-specific QC may be stopped altogether. Individual-patient-type QC may instead be included into the institution's routine QA program in order to continually ensure the machine is capable of delivering the treatment and ensuring the implemented class solutions continue to lead to treatment plans that can be accurately delivered [30].

1.8. Potential Errors in VMAT Treatments

The likelihood of a major error occurring in a radiotherapy treatment that leads to an incorrect dose delivery to the patient is very low. Macklis et al. [31] analysed 93,332 delivered radiotherapy fields and reported an error rate of 0.18%. Margalit et al. [32] reported an even lower error rate of 0.06 % based on the analysis of 241,546 delivered fractions. Nonetheless, even in the last 20 years, there have been a number of major errors in radiotherapy, which are events that caused the patients involved serious injury or death [33 - 39]. Minor errors are events that may lead to slightly suboptimal patient treatments (for example, if the incorrect imaging procedure is used, a patient may be set up incorrectly resulting in incorrect dose delivery to the intended target), and have been shown to occur more regularly in radiotherapy treatment deliveries [40 - 41]. As radiotherapy becomes increasingly complex, our ability to test for errors and catch them if they are present needs to keep pace. This involves determining what errors can occur (a difficult task in itself as new technology may lead to the possibility of new, previously unthought-of errors occurring), including both systematic and random errors. Furthermore, it must be noted that it is not possible to know any measurement with exact precision. Although modern radiotherapy is very precise, there is a limit to how precisely any machine parameters are reported.

1.8.1. Systematic Versus Random Errors

The overall error present in the radiotherapy chain is an accumulation of systematic errors and random errors. Systematic errors are the errors which remain constant or vary in a predictable manner [42].

For example, they may be errors that arise from an incorrect calibration of the treatment machine or may be due to a limitation of the TPS beam model.

Random errors are errors which vary in an unpredictable manner in replicate events. For example, tumour motion during a single fraction is considered a random error as it will vary from fraction to fraction.

Furthermore, errors can be random or systematic for either an individual patient, or a group of patients. For example, an incorrect TPS parameter would be a systematic error that affects a wide group of patients. However, an error in contouring of structures will result in a systematic error for an individual patient but generally constitutes a random error for a group of patients [43]. It is important to consider the risks associated with any such errors. In general, systematic errors will have a net effect that can potentially have a detrimental effect on the treatment, while the impact of random errors is much smaller as they tend to average out over many fractions.

Because systematic errors generally have a larger impact on patient treatments, this study will only investigate systematic errors. Since patient-specific QC is the measurement of an individual patient plan, this study will focus on systematic errors on an individual patient basis as opposed to on a patient population basis (although the potential for the error to systematically effect a patient population will be considered). The sensitivity and specificity of patient-specific QC will be determined by introducing intentional errors and investigating whether the patient-specific QC methods available at the WBCC are able to detect these errors, while assuming that these errors are systematic. It should be noted that some potential (systematic) errors in the radiotherapy chain are not tested by patient-specific QC such as contouring, patient positioning, and intra-fraction motion.

1.9. Accuracy of Patient-Specific QC

Patient-specific QC needs to be sensitive and specific, i.e. reject any treatment plans that contain errors, and pass all treatment plans that do not contain errors. A low sensitivity and specificity will result in errors going unnoticed and acceptable plans being failed. High sensitivity and specificity is not only important from a patient safety perspective, but will also improve the efficiency of patient-specific QC. Sensitivity and specificity have been defined in previous studies [44 - 47], which will be used as a starting point for investigating sensitivity and specificity in this study. These studies also provide a starting point for determining which errors should be introduced, as a number of these studies investigated the effects of introduced errors. Common errors introduced included both errors associated with the linac, such as MLC positioning errors [44 - 45, 48], machine output errors [45, 49], or collimator rotations [49], as well as TPS errors such as varying MLC transmission [50].

1.10. Purpose and Outline of this Study

VMAT was introduced at the Wellington Blood and Cancer Centre (WBCC) in 2012, and it is currently used to treat the majority of patients with head and neck (H&N) and prostate cancers. At the time of commissioning VMAT, a patient-specific QA program was implemented that consisted of three separate methods for conducting patient-specific QC. These are the use of time-resolved point dose measurements [51], Gafchromic film measurements (Ashland Inc., Bridgewater NJ, USA) and array measurements using an ArcCheck device (Sun Nuclear Corp., Melbourne FL, USA). The details of these QC tests are given in chapter 2.

Recently, there has been some discussion in literature questioning the sensitivity and specificity of several patient-specific QC techniques [47, 49, 52]. Furthermore, the sensitivity and specificity of the patient-specific QC methods applied at the WBCC has not been formally quantified. Therefore, the aim of this study is to quantify the sensitivity and specificity of the current patient-specific QC techniques at the WBCC and compare this with literature.

Due to the time constraints for this study, it was not possible to examine every particular radiotherapy error that could potentially occur. As described in section 2.3, a detailed literature review of previous radiotherapy errors and potential errors was carried out to select five error types with potentially the largest impact on treatment outcome. Various magnitudes of these errors were then introduced to a number of VMAT plans which had previously been treated clinically. Error-free treatment plans and plans containing intentional errors were then measured using the three patient-specific QC methods available at the WBCC. A variety of different metrics were used to quantify the sensitivity and specificity of each QC method, and these metrics were compared (when possible) to other previous studies. Finally if the sensitivity and specificity of the current patient specific QC methods was found to be insufficient, recommendations would be made to improve these methods.

A brief overview of each of the following chapters is included below:

Chapter 2 covers the methods and materials used in this study. It starts by outlining the current practise conducted at WBCC in terms of H&N treatment plan generation and patient-specific QC methodology. It then covers the selection of the individual plans that were investigated. The next step was to determine which errors to introduce, starting with defining the clinical relevance of errors in radiotherapy used in this study. A literature investigation into different error modes was conducted and an outline of each error mode to be investigated in this study is presented. How these errors were introduced to each of the original treatment plans is covered and the sensitivity and specificity metrics used for analysis are outlined. Finally, over the course of this study changes to the beam modelling were made, and the methods for undertaking these changes are also outlined in the appendix of this chapter.

Chapter 3 covers the results of this study, starting with the results of the intentional error planning study. The results of each QC method are then highlighted individually. For each QC method, the results of the error-free plan measurements were analysed, then the results of the introduced error plan measurements were analysed using each sensitivity and specificity metric. Finally, the results of the beam model optimisation are included in an appendix to this chapter.

Chapter 4 is a discussion of the major findings from this study. It provides a comparison of the three separate QC methods at the WBCC, and of the sensitivity and specificity of these QC methods to results from other studies. It also highlights the important factors to consider when conducting patient-specific QC such as beam modelling, the clinical relevance of possible errors, and the importance of setting appropriate acceptance criteria for patient-specific QC methods. It also covers some of the limitations of the patient-specific QC methods investigated, as well as some limitations of the analysis methods used throughout this research. Finally, a brief discussion of the benefits of resolving different error modes is given and the recommendations that resulted from this study are stated.

Chapter 5 is the conclusion, which briefly summarises the major findings of this study.

2. Methods and Materials

The purpose of this study was to quantify the sensitivity and specificity of the WBCC patient-specific QC methods, and specifically focus on their ability to correctly identify clinically relevant errors. Nasopharyngeal treatments were selected to be included in this study considering the proximity of the target volumes to multiple organs at risk such as the temporal lobe, brainstem, spinal cord, optic nerve, chiasm, parotid glands, submandibular gland, and the pituitary gland. Due to this close proximity, small errors in the delivery of these plans may result in clinically relevant dose deviations. The standard radiotherapy treatment for nasopharyngeal tumours at the WBCC is VMAT with a prescribed dose of 66 Gy in 30 fractions to the macroscopic disease site utilising two full arcs.

A diagram with an overview of the information flow of the main investigation in this study is presented in **Figure 2.1**. The process was as follows:

1) Initially an error is introduced to a clinical treatment plan, and the plan was measured using a patient-specific QC technique.
2) This measurement was compared to the TPS calculation to give a deviation (Δ) between the measured and calculated values.
3) Δ was then compared to a user defined QC acceptance criteria to determine if the QC result was a positive or a negative result.
4) The result was defined as a true positive, true negative, false positive or false negative (TP, TN, FP and FN respectively) taking into account the clinical relevance of the introduced error.
5) The QC methods were then characterised by calculating the sensitivity (S_1) and specificity (Sp_1) using these results (see section 2.7.1 and 2.8.1) which were subsequently used in a receiver operator characteristic (ROC) curve analysis (see section 2.9).

Figure 2.1 also highlights two branches which correspond to two separate beam models that were used over the course of this study. These are labelled as the 'clinical' beam model (upper branch in **Figure 2.1**) and the 'adjusted' beam model (lower branch in **Figure 2.1**). Over the timeline of this study a new TrueBeam linac (Varian Medical Systems, Palo Alto CA, USA) was installed at the WBCC. During commissioning it was discovered that the measured dosimetric leaf gap (DLG, see section 2.4.4) on the new linac was different to that on the current TrueBeam (which was used for all measurements during this study). Therefore, a sub-project was undertaken to optimise one combined beam model for both linacs. This raised the question of whether the current beam model used clinically (henceforth referred to as the clinical beam model) for the linac in this study was optimised such that TPS calculated data matched measured data as well as possible. A separate investigation was conducted to determine the optimal beam model for this linac alone (the methods and materials for this investigation are described in appendix 2.A and the results are included in appendix 3.A); with the resultant beam model denoted the adjusted beam model.

Further details of the various aspects of the overview in **Figure 2.1**, including the methods and materials applied in this study, are provided in the next sections.

This chapter is organised in the following way:

- Section 2.1 outlines the current practise for VMAT H&N treatment planning, and the selection of treatment plans that were investigated.
- Section 2.2 outlines the definition of clinical relevance that was used for this study.
- Section 2.3 outlines the selection of errors modes that were introduced to the selected treatment plans.
- Section 2.4 describes the error modes that were introduced.

- Section 2.5 details of how each error mode and magnitude was introduced to the selected treatment plans.
- Section 2.6 describes the patient-specific QC methods available at the WBCC and outlines the routine clinical practises used for measurement and analysis of VMAT plans.
- Sections 2.7 to 2.9 define the sensitivity and specificity metrics that were used over the course of this study, as well as the theory and construction of receiver operator characteristic (ROC) curves.
- Appendix 2.A covers the methods used during the beam model adjustment process (appendices are included in chapter 6).
- Appendix 2.B describes the creation of the in-house developed software that was used over the course of this study (appendices are included in chapter 6).

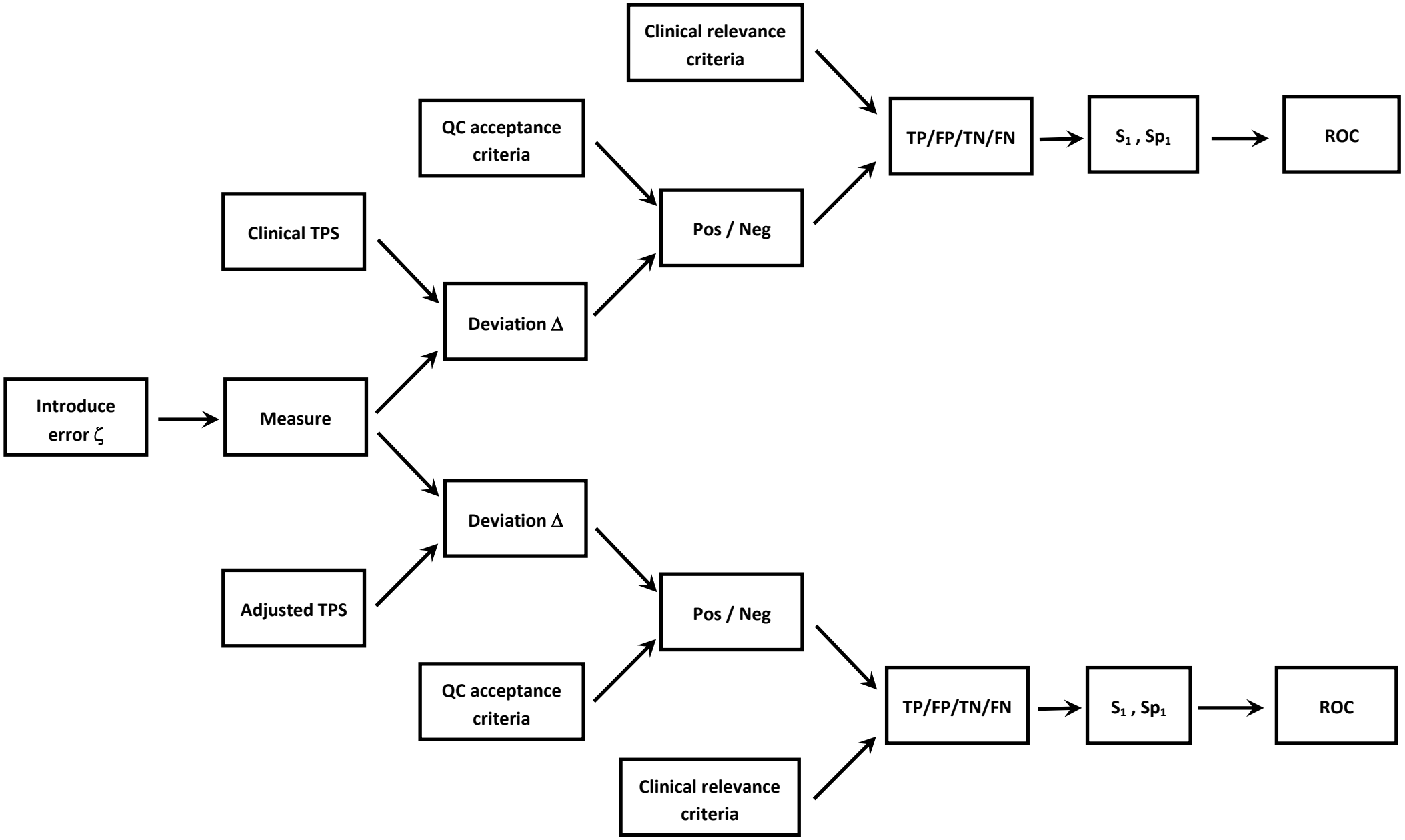

Figure 2.1: *Diagram of the main information flow in this study for introducing delivery errors to clinical treatment plans, measuring them using patient-specific QC, and calculating sensitivity and specificity. ζ corresponds to the type and magnitude of error introduced. Δ is the deviation between the TPS and the patient-specific QC measurement. Pos or Neg refers to whether the QC result passes or fails the acceptance criteria. S_1 and Sp_1 are the calculated sensitivity and specificity respectively (see sections 2.7.1 and 2.8.1).*

2.1. Treatment Planning

2.1.1. Definition of Planning Volumes

All patients requiring treatment for nasopharyngeal tumours underwent CT simulation using a Philips Brilliance BigBore CT scanner (Philips Medical Systems, Eindhoven, The Netherlands). The CT data obtained was fused with Magnetic Resonance (MR) images where available to assist with structure delineation. Target structures and organs at risk were delineated by the radiation oncologist according to WBCC clinical policy. This policy closely follows the consensus guidelines [53], and the International Commission of Radiation Units and Measurements (ICRU) definitions [54]. The gross tumour volume (GTV) is the macroscopic extent of the primary disease, including any involved lymph nodes. The clinical target volume (CTV) with a prescribed dose of 66 Gy (CTV66) was a geometric extension of the GTV that included any possible microscopic spread of disease, but is limited by anatomical borders such as bone and air cavities. In addition, a CTV54 is defined which extends inferiorly down either side of the neck to include lymphatic pathways which may include the microscopic extent of disease, and is prescribed 54 Gy in 30 fractions. A planning target volume (PTV) is an extension of the CTV to account for daily patient set up uncertainties, and to account for patient movement during treatment. The PTV66 and PTV54 were created by expanding the CTV66 and CTV54 by 5 mm in all directions respectively. Additionally, planning PTVs (PPTVs) were contoured for both the PTV66 and PTV54 and these structures were used for plan optimisation. The PPTV66 was defined as the PTV66 with a subtraction of the GTV and cropped 3 mm inside the skin surface. The PPTV54 was defined as the PTV54 cropped 3 mm inside the skin surface, followed by a subtraction of the PTV66, and the resulting structure was cropped so there was a 5 mm margin from the PTV66. The reason these volumes were cropped to 3 mm inside the skin surface was to prevent the optimiser from attempting to deliver dose in the build-up region close to the skin surface.

In addition to the target volumes, organs at risk volumes were contoured by the oncologist and/or radiation therapist. These include the spinal cord, brainstem, parotid glands, eyes, lens', larynx, oral

cavity, mandible, pharyngeal constrictors, optic nerves and optic chiasm. Because the brainstem and spinal cord are often very close to the PTV, a separate planning risk volume (PRV) was created for these structures and was defined as a 5 mm extension from the spinal cord (SC) and brainstem (BS) in all directions (then any slices of these structures that overlap were deleted).

2.1.2. Dose-Volume Histogram Metrics for Clinical Plan Acceptance Criteria

Acceptance criteria for treatment plans at the WBCC are largely based on a number of dose-volume histogram (DVH) metrics. There are four main types of metrics that are relevant to this study.

1) D_x based metrics which specify the minimum dose received by x percentage of the volume.

2) D_{xcc} based metrics which specifies the minimum dose received by x cubic centimetres of the volume.

3) D_{mean} based metrics which indicate the mean dose received by the volume.

4) V_x based metrics which specify the volume (in either cubic centimetres or percentage volume of a structure) that is receiving at least dose x (in either percent or Gy).

2.1.3. Treatment Plan Generation

All plans were created using the Progressive Resolution Optimiser (PRO) [20] module within the Eclipse treatment planning system (Eclipse version 11.0, Varian Medical Systems, Palo Alto CA, USA). In order to optimise a VMAT plan the TPS required that dose volume objectives (DVOs) were defined for the PTVs and OARs. Several treatment machine limitations were also taken into account by the TPS during optimisation (such as gantry rotation speed, MLC leaf travel speed etc.) to ensure the plan can be delivered within the mechanical limits of the linac. The objectives for the outcome of the optimisation process for both target structures and organs at risk are given in **Table 2.1**.

Table 2.1: *DVH objectives for the outcome of the optimisation process for nasopharyngeal cancer treatments at the WBCC.*

Target/Organ	Mean Dose	D_{98}	D_2	Secondary Objectives
PTV 54	54 Gy	95% = 51.3 Gy	107% = 57.78 Gy	-
PTV 66	66 Gy	95% = 62.7 Gy	107% = 70.62 Gy	-
SC PRV	-	-	45 Gy	1 cc ≤ 50 Gy
BS PRV	-	-	50 Gy	D1 = 54 Gy
Chiasm		-	50 Gy	D1 = 54 Gy
Parotid Glands	26 Gy (bilat) D_{50} (ipsilat) < 39 Gy D_{50} (contralat) < 25 Gy	-	-	-

The optimiser initially optimised only a few segments of each arc of the plan which were coarsely spaced throughout the arc. Then as the optimiser increased the resolution level, more segments were optimised until the plan met the departmental planning constraints. A more detailed description of how the optimiser works is outside the scope of this study, but further information can be found in the manufacturer's user manual [55].

The final dose calculation was performed using the Analytical Anisotropic Algorithm (AAA) version 11.0.31 (Varian Medical Systems, Palo Alto CA, USA) with a 1.5 mm calculation grid resolution. For the clinical plans, the dynamic leaf gap was set to 2.0 mm, and the effective target spot size was set to 0.0 mm in both X and Y directions. The MLC leaf transmission was set to 1.35 % for all calculations.

2.1.4. Plan Delivery

All clinical patient treatments were delivered using the 6 MV beam of a Varian TrueBeam linear accelerator (Varian Medical Systems, Palo Alto CA, USA) equipped with a Millennium 120 leaf MLC. The central 40 leaves of each MLC bank had a width of 5 mm, while the outer 20 leaves of each bank had a width of 10 mm at isocentre, allowing a maximum of 40 cm MLC defined fields at isocentre. Maximum leaf speed was set to 2.5 $cm.s^{-1}$, maximum machine dose rate was set to 600

MU.min^{-1} and maximum gantry rotation speed was 6 degrees.s^{-1}. All subsequent measurements made during the course of this study were carried out using the same TrueBeam linac.

2.1.5. Selection of Treatment Plans

Five patients previously treated at the WBCC for nasopharyngeal cancer using VMAT were selected for this study. In particular, treatment plans were chosen where the TPS calculated dose to selected OARs was already close to the critical dose for that OAR. **Table 2.2** lists the disease staging, the volume of the planning target to receive 66 Gy (PTV66), the dose received by 98% of PTV66 (D_{98}), the dose received by 1 cc of the PTV66 (D_{1cc}), as well as the dose to 2% of the spinal cord, brainstem, and optic chiasm for the selected patients (D_2).

Table 2.2: *Summary of the patients included in this study. Disease stage, PTV66 volume, PTV66 D_{98} and D_{1cc}, and D_2 for the BS, SC and chiasm are included. For planning constraints for these volumes, refer to* ***Table 2.1****.*

	Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
Disease Stage	Unknown Primary	Unknown Primary	T3, N1, M0	T4, N1, M0	T1, N0, M0
PTV66 Volume (cm³)	281.5	731.5	59.8	401.9	286.2
PTV66 D_{98} (Gy)	63.0	63.0	64.0	63.8	63.1
PTV66 D_{1cc} (Gy)	69.5	70.0	68.6	70.4	70.8
Brain Stem D_2 (Gy)	47.7	47.5	35.2	62.2	48.4
Spinal Cord D_2 (Gy)	44.4	44.9	39.1	44.5	42.0
Optic Chiasm D_2 (Gy)	4.8	48.6	28.0	60.4	43.4

2.2. Defining Clinical Relevance

An important aim of this study was to determine what magnitude of errors would result in clinically relevant dose deviations to the patient. Therefore it was essential to develop a meaningful definition of clinical relevance. The WBCC protocol for planning VMAT H&N treatments [56] was used as a starting point to achieve a clear definition. This protocol states that the V_{95} must be greater than 98% and that the V_{107} must be less than 1 cc. These constraints are based on the recommendations outlined in ICRU report 83 [54]. Furthermore, the dose to OARs must be below the respective constraint for each critical OAR, such as the spinal cord and brain stem. This constraint is often relaxed for non-critical OARs such as the parotids and oral cavity. The DVH constraints for all targets and OARs are

outlined in **Table 2.1**. These planning acceptance criteria are based on international published guidelines which are founded on historical data on how to achieve high tumour control rates and minimal toxicity rates. The intuitive method to define clinical relevance would be to use the planning acceptance criteria for that purpose. However with this approach, the plan quality (dose conformity, target coverage, and OAR sparing) achieved during plan optimisation of the original plan would have a strong influence on the observed clinical relevance of an introduced error. For example, if a clinical plan had a D_{98} of 99 %, only a large error would result in a D_{98} < 95 % and be considered clinically relevant, whereas if the clinical plan had a D_{98} of 95.1 %, a very small error would be classified as clinically relevant.

The alternative approach is to look at the *change in DVH metrics* caused by introducing each particular error which makes the observed clinical relevance independent of the original plan quality. For the alternative approach, the departmental plan acceptance criteria were combined with commonly accepted requirements for the accuracy of dose delivery. Mijnheer et al. [23] and Thwaites [24] proposed that the delivery of dose to a specified point should be accurate to within ±3.0%. Thwaites stated that a tighter requirement of ≤2.0% would likely be more appropriate to account for systematic uncertainty. Considering this study is primarily concerned with systematic errors, a threshold of ±2.0% was selected as the clinical relevance criterion. However, this alternative approach can only be applied using dose based metrics. Therefore, the volume based metrics in the WBCC planning acceptance criteria were first converted into the equivalent dose based metrics; V_{95} > 98% and V_{107} < 1 cc were re-formulated as D_{98} > 95% and D_{1cc} < 107%, respectively. In this way, the following clinical relevance criteria were derived for this study:

- PTV: ΔD_{98} < -2 %
- PTV: ΔD_{1cc} > +2 %
- OAR: ΔD_2 > +2 % *and* the OAR dose tolerance exceeded

Note that with these definitions, a change in dose is only clinically relevant if the D_{98} *decreases* by more than 2%; or if either the PTV D_{1cc} or OAR D_2 *increases* by more than 2%; **Figure 2.2** displays

an overview of the information flow involved in the definition of the clinical relevance criteria applied in this study.

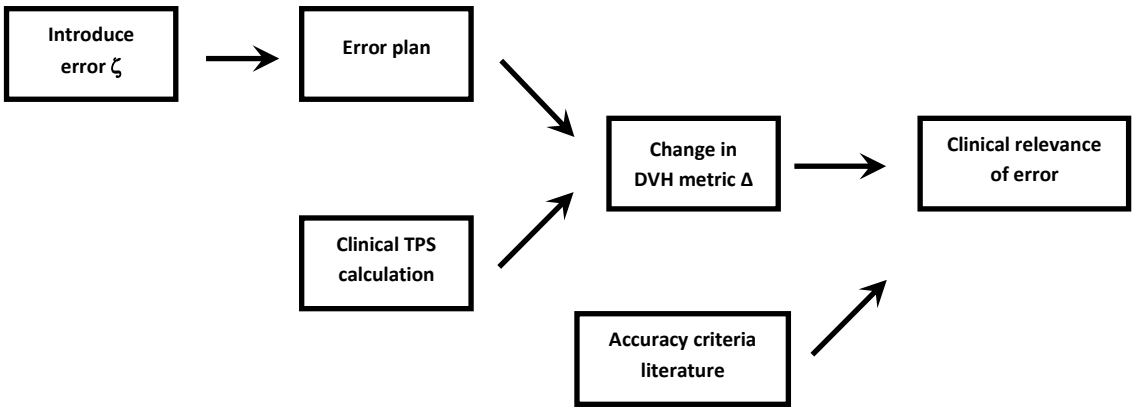

Figure 2.2: *Overview of information flow for determining the clinical relevance criteria. ζ represents the type and magnitude of error introduced. Δ is the deviation between the clinical plan TPS calculation and the introduced error plan TPS calculation.*

As mentioned at the beginning of this chapter, two beam models were utilised for this study. All clinical patient plans used in this study were optimised using the clinical beam model, but were also re-calculated using the adjusted beam model without re-optimisation. Subsequently, values for the DVH metrics mentioned above will be different for the clinical plans depending on which beam model was used to calculate them (see **Figure 2.3**). If an error is then introduced, the change in 3D-location of the isodose lines in the error plan relative to the error-free plan may be the same regardless of which beam model was used for the calculation. However, the effect on a given structure DVH metric (and hence clinical relevance) will generally be different between calculations using either the clinical or adjusted beam models (see **Figure 2.3**). As the re-calculated plans were not optimised using the adjusted beam model, assessment of the clinical relevance of the error using the criteria defined above for these plans does not accurately represent how the error would affect the dose to the target and OAR structures. An example is given in **Table 2.3** for patient 2. For this patient calculation using the clinical beam model, the change in D_{1cc} was above 2% for an introduced 0.5 mm MLC open shift error, and this error was clinically relevant. Whereas a 0.5 mm MLC closed shift did not result in a change in D_{98} over more than 2%. For the adjusted beam model the opposite was true; the 0.5 mm

closed shift was clinically relevant while the 0.5 mm open shift was not. Calculations using the adjusted beam model led to different errors being classified as clinically relevant compared to the clinical beam model. Therefore, the clinical relevance of errors based on calculations using the adjusted beam model is not an appropriate determination of clinical relevance.

Table 2.3: *Subset of the change in PTV DVH metrics for Patient 2.Clinically relevant changes are highlighted in orange.*

Error	Clinical Beam Model		Adjusted Beam Model	
	ΔD_{98}	ΔD_{1cc}	ΔD_{98}	ΔD_{1cc}
MLC 0.5 mm Closed	-1.7%	-1.2%	-2.1%	-1.0%
MLC 0.5 mm Open	1.7%	2.2%	1.9%	1.3%

For this study, the clinical relevance of all errors was exclusively determined using calculations with the clinical beam model. The clinical relevance of all errors was then assumed to be the same over plans calculated using both beam models.

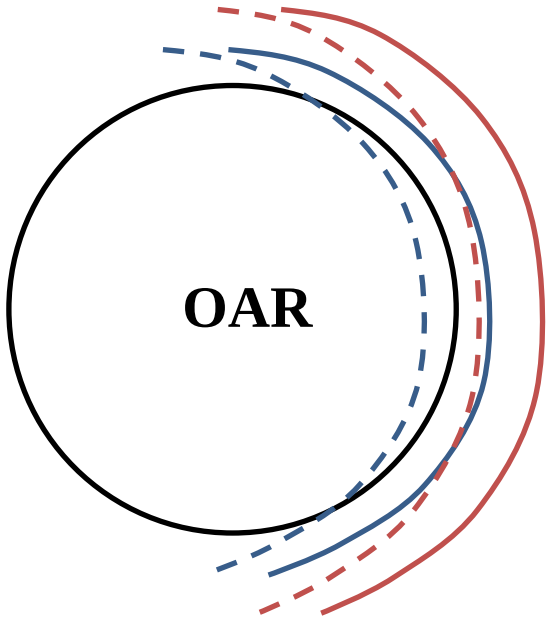

Figure 2.3: *Example of isodose lines around an OAR. Lines in blue corespond to isodoses calculated using the clinical beam model while lines in red were calculated using the adjusted beam model. Dashed lines represent an introduced MLC closed shift error. Solid lines represent the error-free plan calculation. Recalculating the error-free plan using the adjusted model moves the isodoses further away from the OAR. When the closed shift error is applied, it will effect the DVH metrics (as per section 2.2) more for calculations using the clinical beam model (as seen in the overlap of the blue dashed line and the OAR) compared to the adjusted beam model (no overlap between red dashed line and OAR).*

2.3. Selection of Error Modalities

As stated in chapter 1, there is a wide range of errors that could potentially occur in the radiotherapy chain. However, due to time constraints it is not possible to study all of these errors. Therefore, this study will focus on systematic errors on an individual patient level which potentially have the largest impact, and ignore random errors. A literature review was conducted to select the most relevant potential errors modes that can be detected using patient-specific QC and these error modes are listed in **Table 2.4**. For instance, patient set up errors are excluded because they can't be detected using pre-treatment patient-specific QC methods which utilise a phantom (although these errors could potentially be detected using in vivo dosimetry). For each error mode, the likelihood of error occurrence, the probability that the error goes unnoticed and the severity of the error was scored in a similar way to a failure modes and effects analysis (FMEA) using the ranking values outlined in the AAPM TG 100 report [57]. These scores are given in **Table 2.4** and results are presented in a similar manner to a FMEA although a formal FMEA was not carried out. Instead, FMEA-type scoring was applied to the error modes identified from a literature search. Errors which had the highest risk priority number (obtained by multiplying together the scores for likelihood, detectability and severity) would be included in this study.

Where possible, likelihood scores were based on error rates quoted in previous studies. Where no past studies could be found, an estimate of likelihood was made by the author. The likelihood score for errors in TPS parameters takes into account both the possibility of suboptimal values for these parameters being measured, and incorrect data entry of these values in the TPS. Probability that the error goes unnoticed is based on the frequency of conducting routine QC that should detect each error mode. Severity is based on the worst possible patient outcome for the given error mode.

Table 2.4: *Results of the error likelihood, detectability and severity scoring. Errors with the highest risk priority numbers are highlighted in orange.*

Error	Likelihood of occurrence	Probability error goes unnoticed	Severity	Risk Priority Number
MU error	4*	5	9	180
MLC bank error	3**	5	9	135
MLC leaf error e.g. motor error	4	5	4	80
Wrong energy	1	1	10	10
Output variation with gantry angle	3	6	8	144
Gantry angle error	2***	4	5	40
Collimator angle error	2	4	5	40
Incorrect parameter input into TPS e.g. max. dose rate, MLC speed etc.	4	7	2	56
Incorrect MLC leaf modelling e.g. transmission, tongue and groove effect	4	7	3	84
Incorrect DLG modelling in TPS	4	7	6	168
Incorrect focal spot modelling in TPS	4	7	4	112
Small field output error	3	7	4	84

* based on study by Klein et al. [58]
** based on study by Kerns et al. [59]
*** based on study by Margalit et al. [32]

With the approach described above, the error modalities with the highest risk priority numbers were MU errors, MLC bank positioning errors, output variation with gantry angle errors and incorrect DLG or focal spot modelling in the TPS. These five error modes were selected for this study, with a more detailed explanation of each error mode provided in sections 2.4.1 to 2.4.5.

2.4. Error Modes Applied in This Study

2.4.1. Monitor Unit Errors

The output of a linear accelerator is a fundamental property determining the correct delivery of radiation treatment. The output of the linear accelerators is measured using an ionisation chamber (‘the monitor chamber’) in the head of the treatment machine, and is quantified using so-called monitor units (MUs). An MU is defined as a standard unit of dose (units: centigray [cGy]) delivered at the institutes’ reference depth in a water phantom using a standardised measurement

configuration [8]. Therefore, an incorrect number of monitor units (MU) directly represents an incorrect dose delivered to the patient. In particular for VMAT, the machine output with each control point needs to be perfectly synchronised with the change of the other machine parameters during delivery. Therefore it scored highly in the FMEA, and it is important to investigate whether patient-specific QC would detect an MU error.

2.4.2. Output Variation with Gantry Angle Errors

The output of a linear accelerator can vary with gantry angle. One scenario that could lead to this occurring is if the monitor chamber that measures and controls the linac output becomes damaged. In this case the plates of the ionisation chamber might move as the gantry is rotated, which would lead to a systematic variation in output with gantry angle [26]. Furthermore, this error has a high likelihood of not being detected as most linac output calibrations and checks are currently only carried out at one gantry angle, typically with the gantry at the head up position (gantry 0°). This means that a variation in output with gantry angle could have a systematic effect on every VMAT patient until the linac output is measured at non-zero gantry angles.

This study investigated one particular scenario of this error mode. Assuming the linac output is acceptable at its calibration gantry angle (0°), the plates of the ionisation chamber are unlikely to move significantly if rotated through 90° either side of zero. However, when the ionisation chamber's orientation becomes inverted, it is possible for the plates of the chamber to move closer together and therefore cause the linac output to decrease, reaching a maximum deviation at 180°.

It is also worth noting that this error mode would not only effect VMAT treatments, but any radiation treatment where a gantry angle other than gantry 0° is used.

2.4.3. Multileaf Collimator Positioning Errors

During IMRT and VMAT, the aperture of the treatment beam is solely determined by the MLCs. Therefore, an MLC error might lead to overexposure of critical organs at risk. Specifically when it concerns a systematic MLC error, it might occur during multiple treatment fractions in an identical way and result in a dose delivery above the tolerance dose level for that OAR for one or more patients. Furthermore, an MLC error is not completely unlikely as exemplified by an error involving the improper use of an MLC that has been described in literature [60]. In this incident, a series of events lead to the MLC being effectively retracted during an IMRT treatment leading to a clinically relevant overdose to the patient. Although this specific event is unlikely to occur, it would likely have been detected with pre-treatment patient-specific QC. In general, patient-specific QC is most effectively used to detect errors with a potentially large impact for an individual patient. Ideally it will also detect (small) systematic MLC errors that contribute systematically to the overall treatment deviation of many patients.

In the case of a Varian Millennium 120 leaf MLC, the vendor stated tolerance for positional accuracy of each leaf is accurate to within 1.0 mm [61]. Therefore, MLC positional inaccuracies of less than 1.0 mm may routinely occur, which slightly restricts the tolerance for other errors in the overall uncertainty budget. More importantly, potentially larger systematic MLC errors can occur when the MLC is initialised while the linac gantry is not level, leading to the MLC being calibrated incorrectly. These errors could affect each individual leaf, a number of random leaves, or affect either or both MLC leaf banks systematically. Previous studies have indicated that systematic MLC errors can have a large impact on DVH metrics [44, 48]. Therefore this error mode scored highly in the FMEA, and was investigated as a part of this study.

2.4.4. Dosimetric Leaf Gap Modelling Errors

The dosimetric leaf gap (DLG) is a parameter that is used specifically in Varian systems to describe the transmission through the rounded leaf ends of the MLC which effects VMAT, IMRT and static MLC treatments. The leaves of a Varian MLC have a rounded end to ensure similar beam penumbrae at all field sizes. However, as the leaf ends are rounded, the transmission through the leaf ends is noticeable and needs to be accounted for. In Varian systems, this is done by effectively increasing the field size between the leaves by the width of the DLG [62]. It is than important to ensure that the DLG in the planning system and treatment machine are consistent as it has been shown that varying the DLG can affect the beam penumbra and output, and lead to the smoothing of a VMAT dose distribution [62, 63]. This could have the effect of underestimating the PTV coverage (if the DLG is set too small) or increasing the dose to OARs which are in close proximity to the PTV (if the DLG is set too large). It is important to note that the minimum DLG may vary from linac to linac within the manufacturing tolerance of the MLC leaf ends [64]. Because the DLG is a beam modelling variable in the TPS and is set for each energy for each linac, it will systematically affect all VMAT plans using a specific energy. Thus, an incorrect DLG can potentially impact the treatment results for a large group of patients.

2.4.5. Effective Target Spot Size Modelling Errors

For Varian systems, there is a distinction between the focal spot of the treatment machine and the effective target spot size defined in the TPS. The focal spot refers to the physical spot size of the electron beam when it hits the target in the head of the accelerator. The effective target spot size (ETSS) is a parameter in the TPS that can be used to adapt the shape of the geometrical beam penumbra and has an impact on the apparent source occlusion in the TPS [62]. This is particularly the case for small field segments, which are common in VMAT plans.

The ETSS value in the TPS is usually tuned by optimising the match between the measured and TPS calculated beam penumbra. It is possible to adjust the ETSS in both the X (ETSS X) and Y (ETSS Y) directions separately.

It should be noted that in the TPS, these dimensions of the ETSS are related to the collimator rotation with the X direction defined as the direction of leaf travel. In reality, the dimensions of the focal spot of the machine are obviously independent of the collimator rotation. The vendor has presumably chosen for this apparently inconsistent definition in the TPS beam model to limit the number of parameters that can be varied to match the shape of the TPS calculated beam penumbra to the experimentally determined beam penumbra. For instance for a Millennium MLC (Varian Medical Systems, Palo Alto, USA) the penumbra differences in each direction are largely determined by the presence of the rounded MLC leaf tips in the X direction, while the MLC leaves in the Y-direction include tongues and grooves that limit the inter-leaf leakage [65]. The vendor states that a spot size value of 0 mm in both directions may be used, but recommends that the ETSS value in the TPS is optimised for each individual treatment unit [55].

If the default value of 0 mm is used without optimisation, this could introduce systematic errors into the TPS calculations for a large group of patients. For some plans, this systematic error may lead to incorrect dose deliveries at the edge of the PTV, particularly if the PTV and OAR are abutting. Furthermore, the ETSS may be different for various linacs at a particular centre. However, the centre may only use one beam model with an averaged ETSS value in the TPS to simplify use, and this is the current practise at the WBCC. This will also contribute to the total error in the plan calculation.

Similar as with the DLG error mode, an ETSS error will affect every plan calculated (although each plan may be affected differently). Therefore, an error in setting the ETSS could potentially have an impact on the treatment of many patients.

2.5. Error Simulation

There are three separate methods that can be used to investigate the effects of introduced errors on the sensitivity of the various QC methods for the chosen error modes. These are:

1) Send the 'error-free' plan to the linac but adjust parameters on the linac for delivery.
2) Create a new plan with the intentional error and use this to deliver the dose and conduct QC.
3) Simulate the error in the TPS only with opposite sign to that applied for method 2 above, and use this as the 'reference' plan against which to compare the measurement results.

Using method 1, adjusting the individual parameters (e.g. adjusting output out of tolerance, adjusting MLC calibration) on the linac would exactly mimic a real delivery error. However, it is the most labour intensive method as the parameters would need to be adjusted before each different error delivery, and would need to be returned to their clinical settings (and independently verified) at the end of each measurement session. Also this method would be the riskiest as there would be potential for settings to be reset incorrectly, or for other risks such as MLC leaf collisions to occur. Therefore this method of error simulation was not feasible for this study.

Using method 3 to simulate all errors would be the simplest method, as it would involve only making QC measurements of the error-free plans. However this does not necessarily represent how most error modes would occur in reality but assumes the linac and QC system will respond to errors with opposite signs in a similar fashion.

Method 2 involves more work than method 3, but allows delivery and measurement of introduced error plans in a similar way to method 1 without the risk. However, this method cannot be used for error modes which only affect the TPS.

Therefore, both methods 2 and 3 were used to simulate errors in this study. Errors that could affect delivery parameters of the linac (MU errors, MLC positioning errors and output variation with gantry angle errors) were simulated by creating new plans containing intentional errors (method 2 above). Errors that affected the TPS only were simulated in the TPS (DLG and ETSS modelling errors) to calculate verification plans containing the intentional errors, with these plans then being used as the reference plan for the QC analysis (method 3 above).The methods of introducing the errors in this study are described below.

2.5.1. MU errors

Monitor unit errors were introduced by increasing or decreasing the number of monitor units in each arc by a given percentage. The current action level for the daily linac output checks at the WBCC is 3%. Therefore MU increases and decreases of 3% were introduced, as well as a smaller error magnitude of 1.5%. This resulted in four introduced MU error plans per patient.

2.5.2. Output Variation with Gantry Angle Errors

The error-free treatment plans were exported from the Eclipse as DICOM RT plans and were imported into software developed in-house using Matlab (v2014a, The Mathworks, Natick, MA, USA, see appendix 2.B). This software would display the number of MU to be delivered per control point (CP) for each VMAT arc. This software would also allow the user to modify the number of MU delivered per arc segment (see **Figure 2.4**).

This study investigated the output variation with gantry angle error mode described in section 2.4.2. The magnitude of output decreases that were investigated were 4% and 8% at gantry 180°, noting that a 4% decrease is the theoretical maximum deviation possible while still passing routine QC (as it is an international recommendation that the linac output should be within ±2% at any gantry angle [26]).

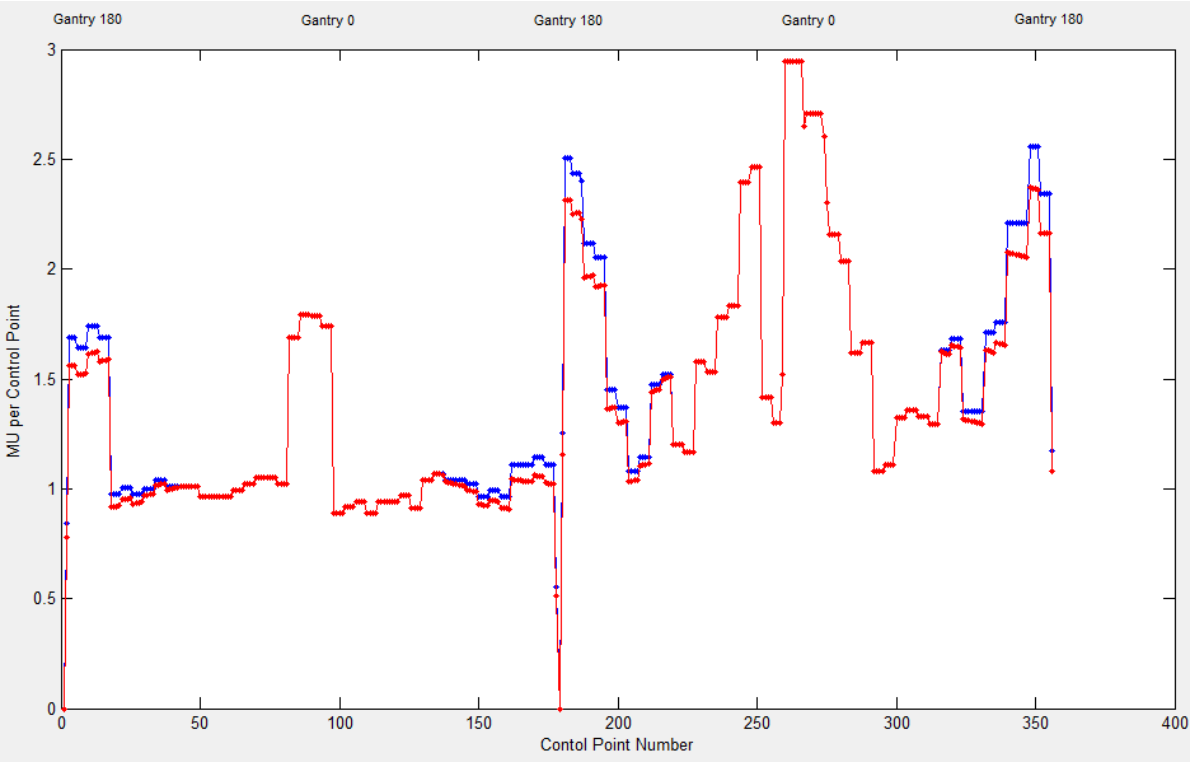

Figure 2.4: *Differential plot of MU delivered per CP for patient 2 showing how the output was modified with gantry angle. The blue line corresponds to the error-free plan delivery, while the red line is the error plan delivery.*

2.5.3. MLC Positioning Errors

MLC errors were introduced by exporting the error-free treatment plans from Eclipse as a DICOM RT file and loading them into separate software developed in house using Matlab (see appendix 2.B). The software application facilitated inspection and modification of the MLC positions using a table, and provided an overview of the MLC settings in a beams eye view (BEV) plot (see **Figure 2.5**).

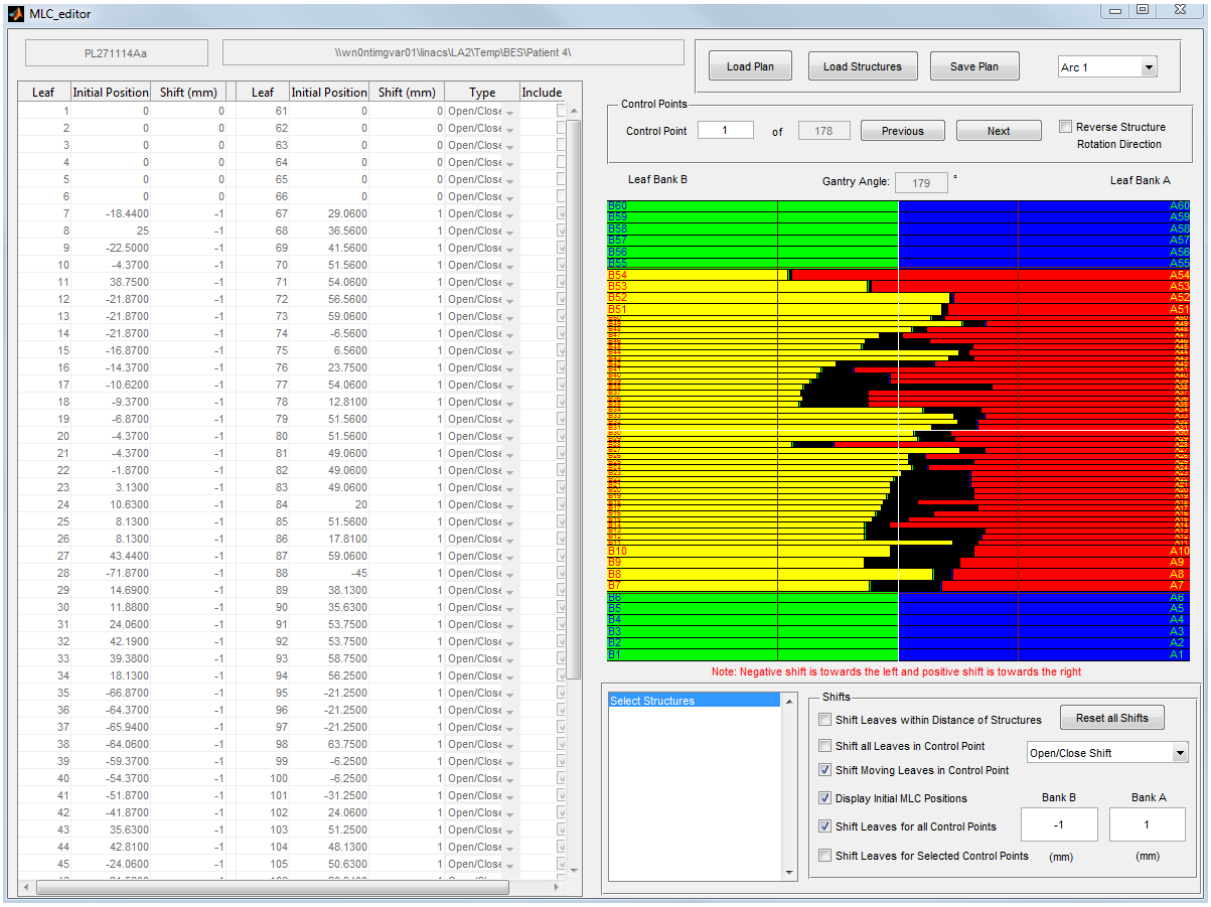

Figure 2.5: *A screenshot of the MLC control software showing the table of MLC leaf positions and the BEV plot.*

Different types of systematic MLC errors were introduced to all moving leaves in each VMAT arc. These included:

- **Closed or open shift errors**: increasing or decreasing distance between opposing MLC leaves by a given magnitude.
- **Translation shift errors**: adding the same offset to the position of both MLC banks.

Restrictions of the MLC positions as set on the linacs were also applied in the software application to prevent leaf collisions or violation of the maximum leaf speed. The application also facilitated reviewing the shielding by the MLCs of contoured structures such as PTVs or critical OARs in a BEV plot (see **Figure 2.6**). In this way, intentional errors could be designed that, at least to a first order approximation, only affected the dose delivery to a single structure. This allowed the ability to

investigate the scenario where an error had a large clinical impact on a given OAR point of interest (POI) at which QC measurements are made as part of routine patient-specific QC at the WBCC.

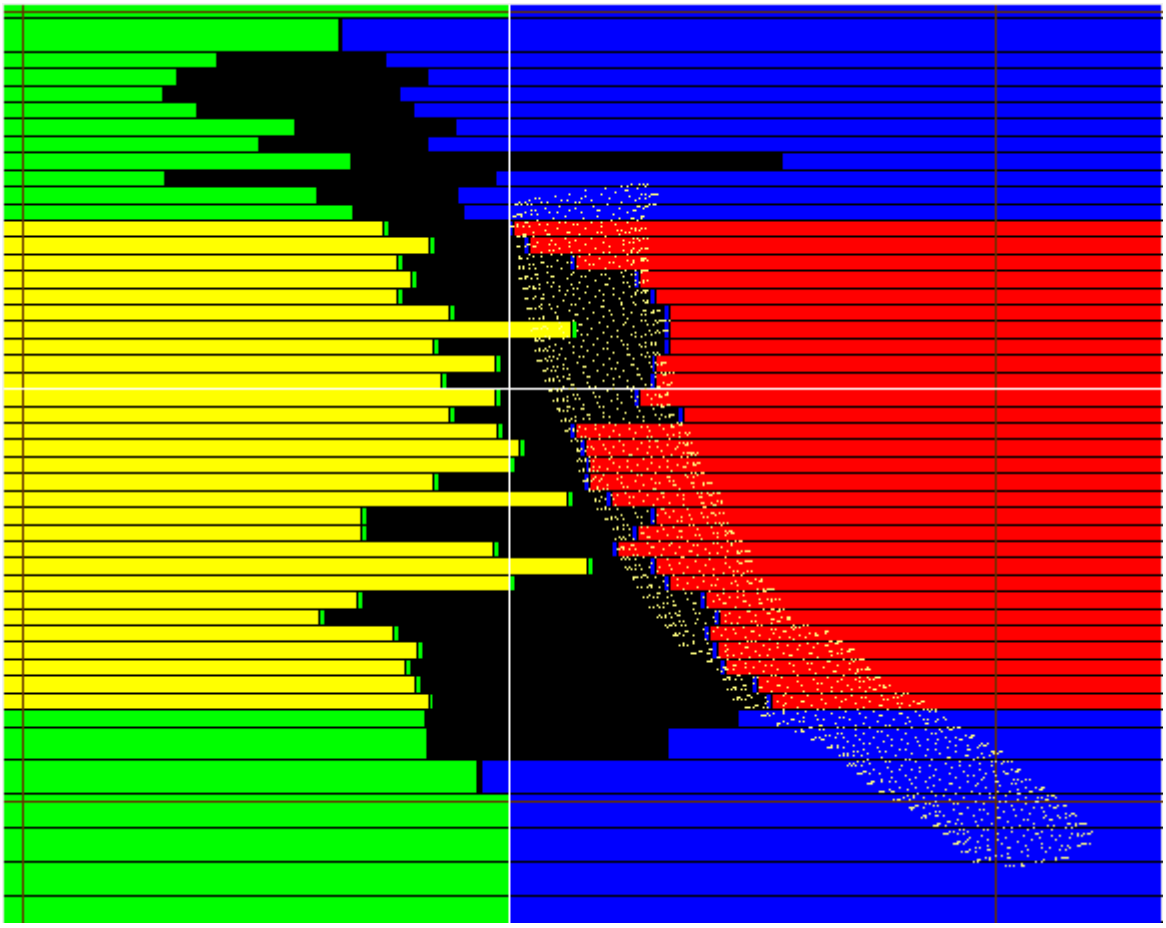

Figure 2.6: *Close up of the MLC display with the spinal cord PRV structure overlaid. Only leaves that are shielding the structure are shifted (although the shift is still applied to both leaf banks).*

Systematic MLC leaf bank shifts

For each error-free patient plan, MLC leaf bank opening and leaf bank closing errors were introduced. A magnitude of 1.0 mm leaf bank shift was chosen, as this is the stated accuracy for the Varian Millennium MLC. This 1.0 mm shift was applied to simulate:

- A 1.0 mm opening of each bank
- A 1.0 mm closing of each bank
- A 1.0 mm translation of both banks in the same direction.

A smaller 0.5 mm shift error magnitude was also investigated for open and closed MLC shifts errors only.

Specific OAR MLC leaf errors

Four MLC errors relating to specific OARs were introduced. For each patient, 3 separate plans were produced with a 2.0 mm open shift error applied to any MLC leaf shielding the SC PRV, BS PRV and optic chiasm respectively. One plan was also produced which contained a 1.0 mm open shift error applied to any MLC leaf shielding the SC PRV for each patient. **Table 2.5** provides a summary of MLC positioning errors introduced for each error-free patient plan.

Table 2.5: *Summary of MLC errors introduced for each error-free patient plan. The measurements points/planes are also indicated for each error.*

Type of MLC shift	Magnitude of MLC shift	Leaves involved
Open	0.5 mm	All moving leaves
Open	1.0 mm	All moving leaves
Closed	0.5 mm	All moving leaves
Closed	1.0 mm	All moving leaves
Translation	1.0 mm	All moving leaves
Open	2.0 mm	Leaves overlapping SC PRV
Open	1.0 mm	Leaves overlapping SC PRV
Open	2.0 mm	Leaves overlapping BS PRV
Open	2.0 mm	Leaves overlapping chiasm

2.5.4. Beam Modelling Errors

As mentioned in the introduction to this chapter, two separate beam models were implemented during this study, the clinical beam model and the adjusted beam model. Only two parameters of these models differed: the DLG and the ETSS in the X direction. In addition to investigating the effect of using different values for both these parameters at once, the effect of a different DLG or ETSS X alone was also investigated.

2.5.5. DLG Modelling Errors

A virtual linac was set up in the TPS which contained identical beam data to the physical linac but allowed for a different DLG to be used. Additional verification plans were generated from the error-free patient plans, with the VMAT beams converted from the physical treatment linac to the virtual linac with the modified DLG. These verification plans were then recalculated using the same AAA dose calculation algorithm, and were used as the TPS reference dose in the QA techniques for comparison with the measurement results of the error-free patient plans.

Two DLG values were investigated as a part of this study:

- A DLG value of 2.0 mm which was the setting used clinically when the treatment plans being investigated were created.
- A DLG value of 1.2 mm which was found to provide the optimal fit with measured data (see Appendices 2.A and 3.A).

When simulating this error mode using the adjusted beam model, the DLG matches the correct value, but the ETSS X value is different from the correct value (see section 2.5.6). Therefore this error becomes the ETSS X error (but with a negative sign i.e. a -1.5 mm ETSS X change) when the adjusted beam model is used.

2.5.6. ETSS Modelling Errors

Since the effective target spot is a parameter that only exists in Eclipse to optimise the match between the measured beam penumbra and the TPS calculated beam penumbra, any error of this type cannot be physically measured on the linac. Therefore, a similar approach was used as for the introduction of DLG errors (see section 2.5.5).

In Eclipse, the ETSS is linked to the selected calculation algorithm for dose calculation. Therefore additional AAA algorithms were created that were identical to the clinically used AAA algorithm except for the values of the ETSS to simulate ETSS errors. In this study, only the dimension of the target spot in the x direction was modified, as the ETSS in the y direction was the same in the clinical beam model and the adjusted beam model (see appendix 3.A), and beam modelling errors were only introduced by changing parameters from their clinical model settings to their adjusted model settings.

Additional verification plans were generated from the error-free patient plans, with the dose determined using the AAA algorithm with the modified ETSS parameters. These verification plans were then used as the TPS reference dose in the QA techniques for comparison with the measurement results of the error-free patient plans.

Two ETSS X values were investigated as a part of this study:

- An ETSS X value of 0.0 mm which was the setting used clinically when the treatment plans being investigated were created.
- An ETSS X value of 1.5 mm which was found to provide the optimal fit with measured data (see Appendices 2.A and 3.A)

When simulating this error mode using the adjusted beam model, the ETSS X matches the correct value, but the DLG value is different from the correct value (see section 2.5.5). Therefore this error becomes the DLG error (but with the opposite sign i.e. a +0.8mm DLG change) when the adjusted beam model is used.

2.5.7. Error Simulation – Multiple Errors per Plan

A plan was also produced which included two introduced errors for each patient. This would facilitate the investigation whether the QC methods could resolve various types of error modes within a single measurement. For this plan, the method of introducing these errors was the same as that outlined in the above sections. This plan contained two introduced delivery errors (a 3% MU increase and a 1.0 mm MLC closed shift) to determine how the sensitivity of each QC method is affected when two error modes of opposite magnitude are applied to one plan.

Furthermore, all plans with a single error introduced (as well at the error-free plan) were calculated using two different beam models (see appendices 2.A and 3.A for details). This would facilitate investigation into how the sensitivity of each QC method varies through changing beam modelling parameters.

2.5.8. Summary of Plans with Introduced Errors

A summary of all plans with introduced errors is included in **Table 2.6**. These are split into plans with delivery errors (output errors, MLC positioning errors and output with gantry rotation errors) and TPS errors (DLG errors and ETSS X errors). Overall, 18 plans containing introduced errors were produced for each of the five error-free patient plans. There was one exception for patient 1; in this plan the chiasm not located in the treatment field for any segment of either arc, therefore the 2.0 mm open shift error near the chiasm was not implemented for this patient. Overall 89 plans containing introduced errors were produced for this study. Additionally, all plans with delivery errors were calculated using two separate beam models (see appendices 2.A and 3.A for details resulting in 79 introduced error plans (as well as the 5 error-free plans) for comparison across both beam models.

Table 2.6: *List of errors to be introduced for each patient. Note that green cell shading in the delivery and TPS parameters columns indicates the correct value for that given parameters, while red cell shading indicates an introduced error in that given parameter.*

Plan error types	Delivery Parameters		TPS Parameters	
	MU	MLC	DLG	ETSS X
Error-free plan	Correct	Correct	2.0 mm	0.0 mm
Output 1.5 % decrease	1.5% low	Correct	2.0 mm	0.0 mm
Output 3 % decrease	3 % low	Correct	2.0 mm	0.0 mm
Output 1.5 % increase	1.5% high	Correct	2.0 mm	0.0 mm
Output 3 %increase	3 % high	Correct	2.0 mm	0.0 mm
Output with gantry angle 4%	4% low @ G180°	Correct	2.0 mm	0.0 mm
Output with gantry angle 8%	8% low @ G180°	Correct	2.0 mm	0.0 mm
MLC closed 1 mm	Correct	1.0 mm closed	2.0 mm	0.0 mm
MLC closed 0.5 mm	Correct	0.5 mm closed	2.0 mm	0.0 mm
MLC open 0.5 mm	Correct	0.5 mm open	2.0 mm	0.0 mm
MLC open 1 mm	Correct	1.0 mm open	2.0 mm	0.0 mm
MLC Translation 1mm	Correct	1.0 mm Translation	2.0 mm	0.0 mm
MLC SC 1.0 mm	Correct	1.0 mm open near SC	2.0 mm	0.0 mm
MLC SC 2.0 mm	Correct	2.0 mm open near SC	2.0 mm	0.0 mm
MLC BS 2.0 mm	Correct	2.0 mm open near BS	2.0 mm	0.0 mm
MLC chiasm 2.0 mm	Correct	2.0 mm open near Chiasm	2.0 mm	0.0 mm
MU 3% increase, MLC closed 1 mm	3 % high	1.0 mm closed	2.0 mm	0.0 mm
DLG 1.2 mm	Correct	Correct	1.2 mm	0.0 mm
ETSS X 1.5 mm	Correct	Correct	2.0 mm	1.5 mm

2.6. Patient-Specific QC Methodology

All error-free plans and plans containing introduced errors underwent patient-specific QC using the methods outlined in the following sections:

2.6.1. TPS Dose Calculation

The plan dose needed to be calculated by the TPS for comparison against the measured dose for all QC methods. Both the trPD and film QC methods utilise the same initial verification plan dose calculation. The TPS dose was calculated by producing a verification plan of the clinical patient plan. This was done by taking the clinical VMAT plan and copying it onto a CT data set of the plastic water phantom (see **Figure 2.7**) and then recalculating the plan dose using the appropriate beam model. The Hounsfield Unit (HU) values for the plastic water phantom were overridden to 0 HU. A virtual couch

structure was used in Eclipse to simulate the attenuation through the couch with the couch surface given a value of -450 HU and the body of the couch had a value of -970 HU at the WBCC. Measurement points were selected by creating reference points in the verification plan that geometrically corresponded to points of interest in the clinical patient plan. The TPS dose was then calculated using the same AAA dose calculation algorithm used to calculate the clinical plan, using either of the two beam models applied in this study, and a 1.5 mm dose grid resolution. The treatment plan and dose calculation were then exported as DICOM RT files ready to be imported into the respective analysis software for trPD and film QC.

For the trPD QC analysis, the TPS dose per segment is required. However, Eclipse did not provide this data and a work around had to be introduced. The verification plan was exported to in-house software that converted each of the VMAT beams into a new plan consisting of 177 static beams based on the 177 segments of the VMAT beams. The MLC positions of each static beam were defined as the average of the MLC positions defined at the CPs either side of the segment, while the MU weight of each static beam was defined as the MU delivered between the two CPs at either side of the segment. The two converted plans (one for each VMAT beam) were then reimported into Eclipse and the dose was recalculated using the same algorithm and dose grid resolution. All verification plans were calculated using both the clinical beam model and the adjusted beam model. After calculation, the plans were then exported from the TPS to the in-house software to perform the time resolved analysis.

TPS dose calculation for the ArcCheck was conducted separately and the details of this are given in section 2.6.4.

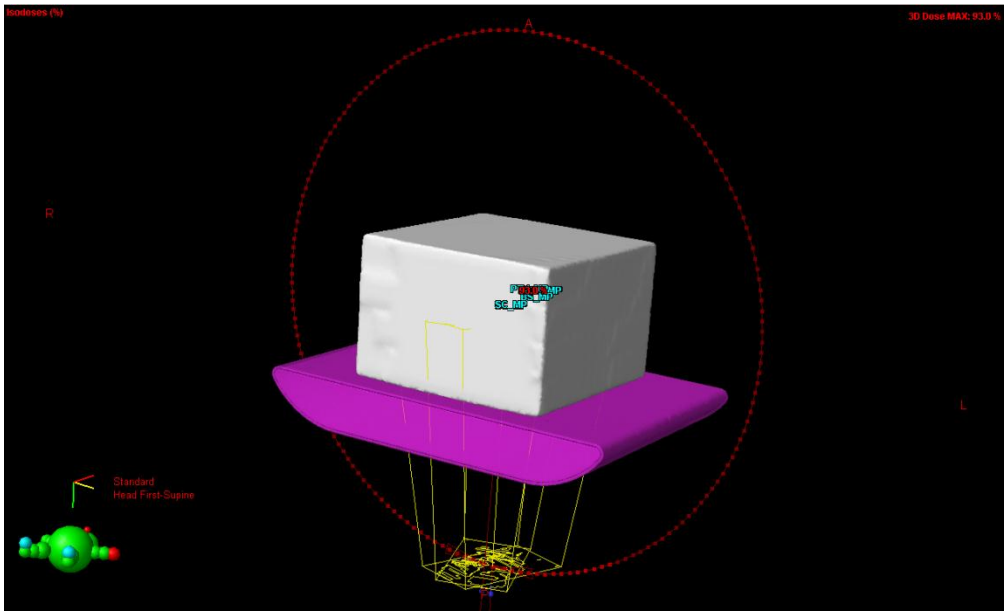

Figure 2.7: *CT data set used to calculate the verification plan dose for point dose and film measurements.*

2.6.2.Time Resolved Point Dose (trPD) Measurements

Equipment

Time resolved point dose measurements [51] were conducted using a 0.015 cc PTW 31014 pinpoint ionisation chamber (PTW, Freiburg, Germany) [66]. This ionisation chamber was cross-calibrated against the WBCC local reference ionisation chamber (PTW 30012 Farmer type chamber) in a 6 MV photon beam in accordance with the TRS-398 code of practise [67].

All measurements were made in a 30x30x20 cm^3 plastic water slab phantom ("The Original", Computerized Imaging Reference Systems, Norfolk, VA, USA). The slab phantom consisted of individual plastic water slabs ranging from 0.1 cm to 5 cm in thickness. One 2 cm thick slab of plastic water had a cavity drilled into it in which the pinpoint ionisation chamber was inserted for measurements. This cavity was drilled such that the sensitive volume of the ionisation chamber was located directly at the centre of the slab. The order of the plastic water slabs could be adjusted such that the chamber could be placed at any height in the phantom, and the phantom could be moved to facilitate measurements at various positions longitudinally. However, the chamber slab only had a

cavity at one position which could not be moved laterally, and the phantom was always positioned isocentrically. Therefore, measurement points were limited to locations within the sagittal plane positioned at isocentre.

Verification Plan Delivery

Error-free plans and plans including intentional errors were delivered using a TrueBeam linac. Current WBCC practise is to conduct trPD measurements at a POI that represents the centre of the PTV. This is because this point is in a high dose region with a low dose gradient and it has been shown that the accuracy of trPD measurements is reduced in POIs where there is a high dose gradient or a low total dose [51]. For MLC shift errors near specific OARs, trPD measurements were also made at the centre of the specific OAR as well as the PTV point to investigate if trPD measurements at the PTV location were as sensitive as trPD measurements at the given OAR location.

Time Resolved Signal Acquisition

The pinpoint ionisation chamber was connected to a PTW T10016 Tandem electrometer which provided a polarising voltage of -300 V. Prior to each measurement session the ionisation chamber was left connected to the electrometer for at least 10 minutes to allow stabilisation, and was then pre-irradiated with 600 MU to reduce leakage current. The electrometer was operated using in-house developed software [68] and time resolved data was obtained by operating the electrometer in continuous streaming mode with a readout frequency ($f_{readout}$) of 10 Hz.

The electrometer reported average current over the 0.1 s readout period, therefore the accumulated charge (ΔQ_t) during each read out period was be obtained using Equation 2.1:

$$\Delta Q_t = I_{t,ave} \times \Delta t \qquad \text{Equation 2.1}$$

Where:

- $I_{t,ave}$ is mean measured current over readout period
- Δt is the readout period which is equivalent to $f_{readout}^{-1}$

For each measurement, additional data was obtained for at least 60 s prior to turning the first radiation beam on, and for 60 s after termination of the final radiation beam of a treatment plan. This extra data was obtained to correct for both pre and post irradiation leakage. The in-house software saved the measurement data into a comma separated values (CSV) file that was subsequently imported into the analysis software.

Time Resolved Analysis

Pre-irradiation leakage was approximated as a linear model fitted to the 60 s of data acquired before the beam delivery began, and this was subsequently subtracted from all time resolved data. Post irradiation leakage was approximated by fitting the 60 s of data acquired after the termination of the radiation beam to an exponential decay model, and this was then subtracted from the time resolved data [51]. The measured dose per time interval, $D_{t,meas.}$ is calculated using Equation 2.2:

$$D_{t,meas.} = Q_{t,corr}.N_{d,w,q0}.K_{q,q0}.K_S.K_{pol}.C_w.K_{T,p}.O \qquad \text{Equation 2.2}$$

Where:

$Q_{t,corr}$	=	the corrected charge reading for time, t,
$N_{d,w,q0}$	=	the dose to water calibration factor for the pinpoint ionisation chamber in a Co-60 beam determined by in-house cross calibration,
$K_{q,q0}$	=	the beam quality correction factor to convert from Co-60 to the 6 MV linac beam used in measurement,
K_s	=	the ion recombination factor,
K_{pol}	=	the polarity correction factor,
C_W	=	the plastic water to water correction factor,
$K_{T,p}$	=	the temperature and pressure correction factor, and
O	=	the linac output correction factor.

The linac log files were then used to determine the relative timing of each segment within each arc during plan delivery. The analysis software application automatically determined the absolute timing of each arc maximising the cross-correlation signal between the experimental and calculated data while varying the time offset of each arc separately. If necessary, the absolute timing of each arc could be adjusted manually [51]. The analysis software facilitates various analysis options. The initial analysis used a plot of $D_{t,meas.}$ as a function of CP number to verify the synchronisation of the measured data with the TPS calculated data. Furthermore, the delivered dose per CP could be plotted

as a function a various other parameters such as gantry angle, field size and position of the detector relative to the MLC defined field edge. An example of a time-resolved analysis showing the dose delivered per CP from both the measured and calculated plans is given in **Figure 2.8**.

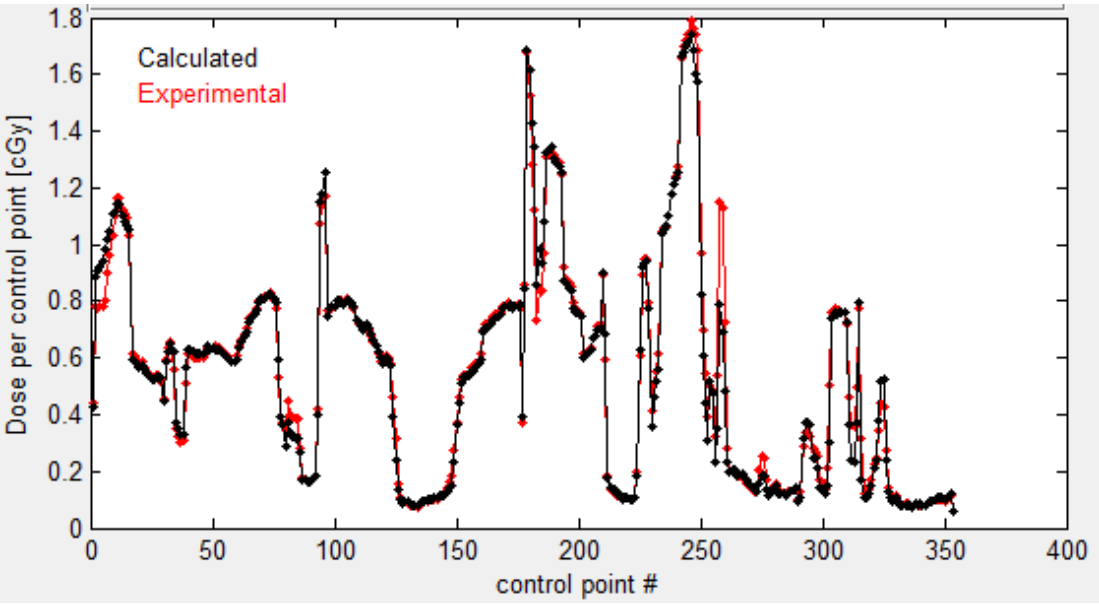

Figure 2.8: *Example time resolved point dose analysis, showing dose delivered per CP for the measured data (Red) and TPS calculated data (Black).*

The software developed in-house using Matlab allows various ways to analyse the time-resolved results [51]. However, only two methods were used here; the integral fractional dose difference and the dose deviation per CP against the detector distance to field edge (DTFE).

Integral Fraction Dose Difference

For our departmental routine QC the integral fraction dose difference (ΔD_{int}) was used to apply an acceptance criterion. This was calculated using Equation 2.3:

$$\Delta D_{int} = \frac{(D_{exp} - D_{TPS})}{D_{TPS}} \qquad \text{Equation 2.3}$$

Where:

- D_{exp} is the integral experimental dose obtained by summing $D_{t,\,meas.}$ for every segment
- D_{TPS} is the dose calculated by the TPS to the measurement point in the verification plan

The current WBCC acceptance criterion for point dose measurements is:

$$|\Delta D_{int}| \leq 2.0\%$$

Dose deviation per control point against DTFE

The position of the detector in the phantom is known, along with the gantry, collimator and couch angles which are defined in the treatment plan. Therefore position of the MLC leaves and detector location in the linac beams eye view (BEV) can be determined (see **Figure 2.9**) for each CP. From this BEV plot, the distance from the detector reference point to the nearest MLC edge can be calculated for each CP of the plan.

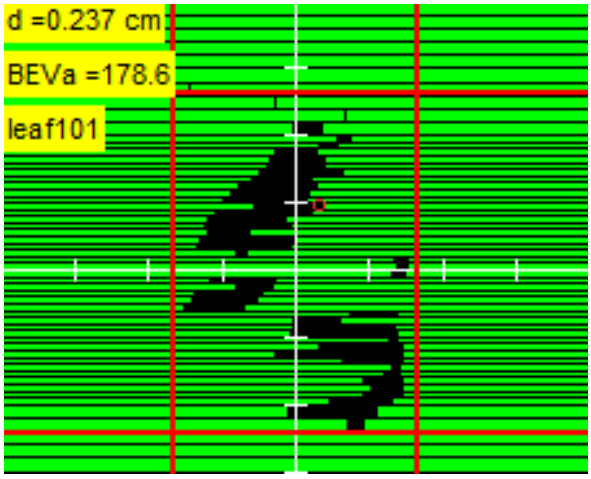

Figure 2.9: *Beam's eye view (BEV) image of one CP, showing the MLC pattern (green leaves), jaw positions (red lines), linac cross hairs (white lines) and detector reference point (red circle). The distance to the nearest MLC leaf is also shown (0.237cm), along with the BEV area, and the number of the closest leaf (leaf 101).*

This DTFE position is averaged over two adjacent CPs to get the average distance from the field edge over that segment, and this is plotted against the difference between measured and calculated dose over that segment (see **Figure 2.10**).

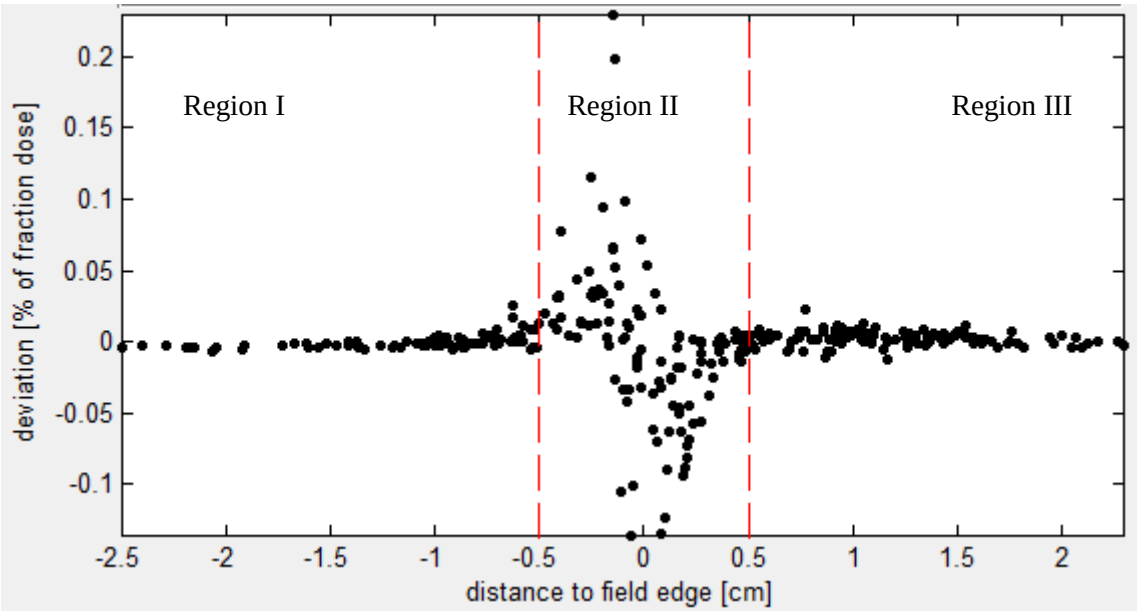

Figure 2.10: *Plot of dose deviation against distance to field edge (DTFE) for a single plan verification measurement. Red lines indicate 0.5 cm from the MLC defined field edge in each direction.*

A plot of dose deviation against DTFE can be broken up into three sections (see **Figure 2.10**);

Region I: $DTFE < -DTFEcrit$
Region II: $-DTFE_{crit} \leq DTFE \leq DTFE_{crit}$
Region III: $DTFE > DTFEcrit$

Where $DTFE_{crit}$ is the threshold which defines the regions where the detector is behind the MLC leaves / near the beam penumbra / in the open field (in **Figure 2.10** above, $DTFE_{crit}$ is set to 0.5 cm). This analysis was utilised to investigate the ability of the trPD method to resolve different error modes (see section 2.8.2).

2.6.3. EBT3 Gafchromic Film Measurements

Equipment

EBT3 gafchromic film (Ashland Inc., Bridgewater NJ, USA) was used to measure 2D dose distributions. In order to allow direct comparison of film results against trPD results, EBT3 film dosimetry was carried out using the same plastic water slab phantom as for time-resolved point dose

measurements. The films were aligned isocentrically in any coronal plane within the slab phantom, allowing measurement at any coronal plane.

Error-free plans and plans containing introduced errors were delivered using a TrueBeam linac. Current WBCC practise is to conduct film measurements at the same coronal plane that includes the corresponding trPD measurements. Therefore films were irradiated at the plane corresponding to the PTV location for all error-free plans and plans containing introduced errors. For MLC shift errors near specific OARs, films were also located at the plane corresponding to the given OAR to investigate if film measurements in the PTV plane were as sensitive as film measurements at the given OAR plane.

Films were scanned at 72 DPI in transmission mode using an Epson 10000XL scanner and were analysed using software developed in-house using Matlab (see appendix 2.B). During each scan session, an empty scan (no film) and a blank film scan (non-irradiated film) were included to enable calculation of the net optical density (OD).

Film Scanning Protocol

Films were left for at least 18 hours before being scanned to reduce the variation in post irradiation darkening [69]. The non-irradiated and irradiated films were handled with gloves at all times and the plastic water blocks were cleaned prior to placing the film on them to avoid contamination. The films were also kept in low light conditions to reduce darkening due to light exposure.

A key component of film dosimetry is the reproducibility of the scanning protocol to reduce variation in the scanner response [70]. For each set of exposed films the following scanning procedure was used:

- Scanner was warmed up with 20 scans performed consecutively in quick succession.

- A frame was used to position the films reproducibly in the scanner (see **Figure 2.11**).
- Each film was scanned four times in quick succession to overcome the cooling down effect of the scanner after opening the scanner lid between films. The fourth scan of each film was then used for analysis to minimise the variation of scan conditions.
- Each set of irradiated films was scanned during two separate scan sessions to assess the scanner reproducibility on different days.

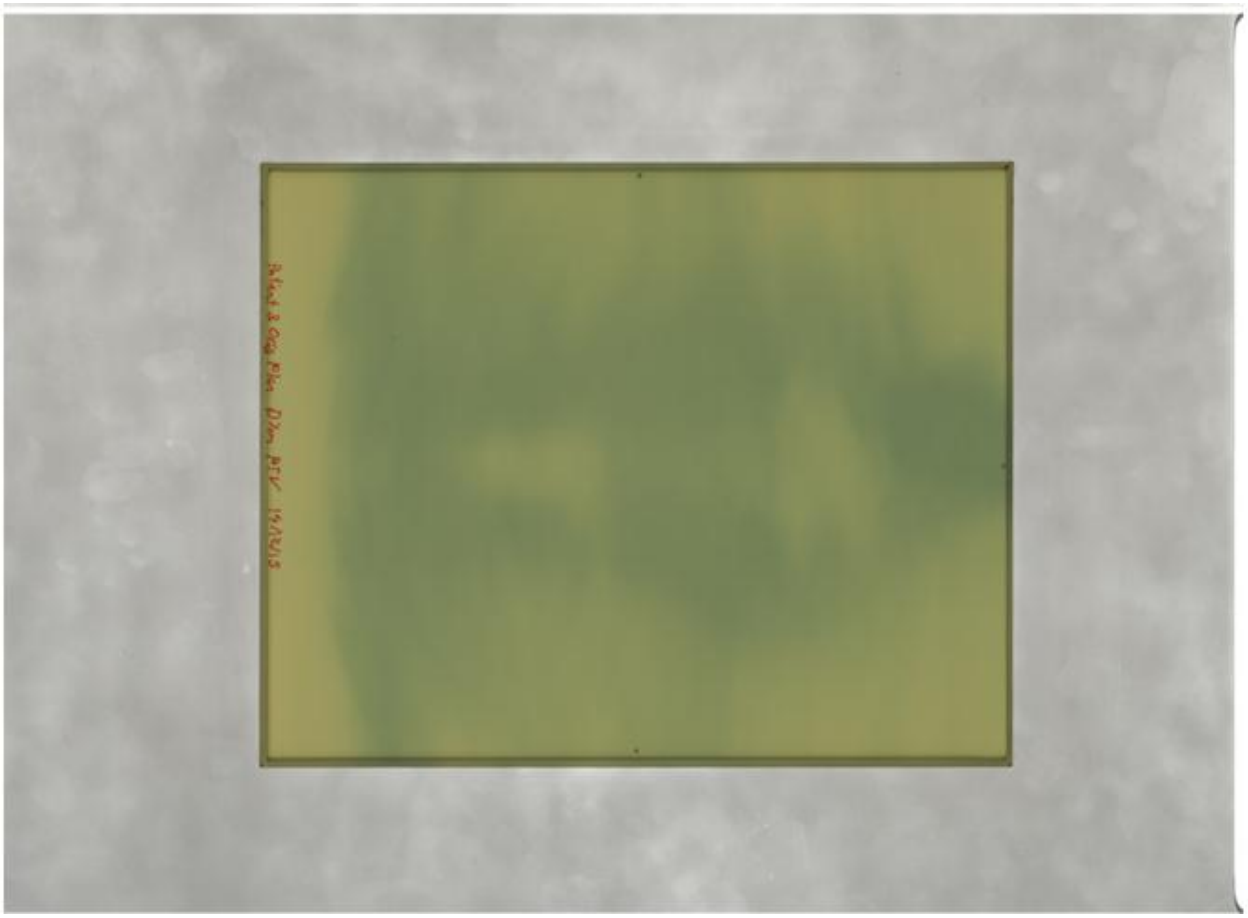

Figure 2.11: *Example of a typical irradiated EBT3 film scan. The grey holder is used to reproduce the position of each film to be scanned.*

Optical Density Calculation

The optical density of the film (OD_{film}) is calculated using one of the three colour channels - red, green or blue and the raw pixel values from the scanned films using Equation 2.4:

$$OD^{c}_{film} = -\log\left(\frac{PV^{c}_{film}}{PV^{c}_{empty}}\right)$$ Equation 2.4

Where:

- PV_{film} is the pixel value of the scanned film
- PV_{empty} is the pixel value of the empty scan
- c is the colour channel

To determine the OD change induced by irradiation only, the net OD (OD_{net}) of an experimental film was calculated using the red colour channel and Equation 2.5:

$$OD_{net} = \frac{OD^{r}_{film}}{OD^{b}_{film}} - \frac{OD^{r}_{blank}}{OD^{b}_{blank}}$$ Equation 2.5

Where:

- OD_{blank} is the optical density of a non-irradiated (blank) film

As the dose dependence of the blue channel is at least one order of magnitude smaller compared to the red channel, the vendor advises to use the information of the blue channel to correct for film thickness variation which reduces uncertainty resulting from the variation in the thickness of the individual films. Therefore, the variation in thickness of each film was corrected by dividing the red channel OD by the blue channel OD. The net OD was subsequently used for dose calibration.

Sensitometric Curve

A jaw defined step-wedge plan that included ten different dose levels from about 0.0 Gy to about 3.9 Gy was created in Eclipse to determine the sensitometric curve. This plan was used to irradiate a calibration film with the film at 5 cm depth in plastic water and 95 cm SSD, with the wedge direction along the superior/inferior axis of the film (see **Figure 2.12**). Considering the limited accuracy of the TPS in calculating the delivered dose in tails of a profile, an error of up to several percent can be expected at the lowest dose levels of the step wedge due to accumulation of this error. Therefore, the

delivered dose at each dose level was measured using an ionisation chamber, and a separate plan was created in the TPS to simulate the actual dose delivered to the calibration film to within 0.2%.

Figure 2.12: *An example of a step-wedge calibration film. The ten bands correspond to ten separate dose levels ranging from 0.0 Gy to 3.9 Gy.*

The sensitometric curve that relates dose to net OD was modelled using a gamma distributed single hit model [71] as given in Equation 2.6:

$$D(OD^{net}) = ae^{\frac{-\log(1-\frac{OD^{net}}{c})}{b}} \qquad \text{Equation 2.6}$$

Where:

- $D(OD)$ is the dose for a specific optical density
- a, b and c are the curve fitting parameters

In addition, the lateral scan effect of the scanner needs to be corrected for [70]. Therefore, an individual sensitometric curve was determined at each lateral position *i* using Equation 2.7:

$$D(OD^{net})_i = a_i e^{\frac{-\log(1-\frac{OD_i^{net}}{c_i})}{b_i}}$$ Equation 2.7

The lateral dependence of the fitting parameters (a_i, b_i and c_i) was approximated by a second degree polynomial. A global fit was carried out to optimise the correspondence between the measured and reference calibration dose using the Levenberg-Marquardt least-squares curve fitting algorithm [72] [73]. These curve parameters were then saved into a calibration file and were used to convert the net OD to dose for all experimental films of the same measurement session.

Analysis

The scanned measurement film, blank film, empty scan and calibration data set were imported into the in-house Matlab software along with the TPS plan and dose matrix. A screenshot of the EBT film analysis software is shown in **Figure 2.13**. The measurement film was manually registered by the user to accurately align it to the TPS dose to within 1 mm by minimising the dose differences in the global dose map (see bottom right panel of **Figure 2.13**). The measured dose was corrected for the linac output during the film measurement session. Comparison of film planes with the corresponding TPS dose plane was conducted using global gamma analysis [74] using a gamma (γ) evaluation criterion of {2%, 2 mm}. ΔD values were calculated relative to the average dose delivered to the high dose area of the film (where the average PTV dose is determined as 80% of the 99th percentile film dose) that approximated the PTV in the dose plane. A threshold of 50% of the maximum plane dose was used for γ analysis. The current WBCC acceptance criteria for film QC are:

i. $\{2\%, 2\ \text{mm}\}\ \gamma\ \text{pass rate} > 85\%$
ii. $\gamma_{ave} < 0.5$

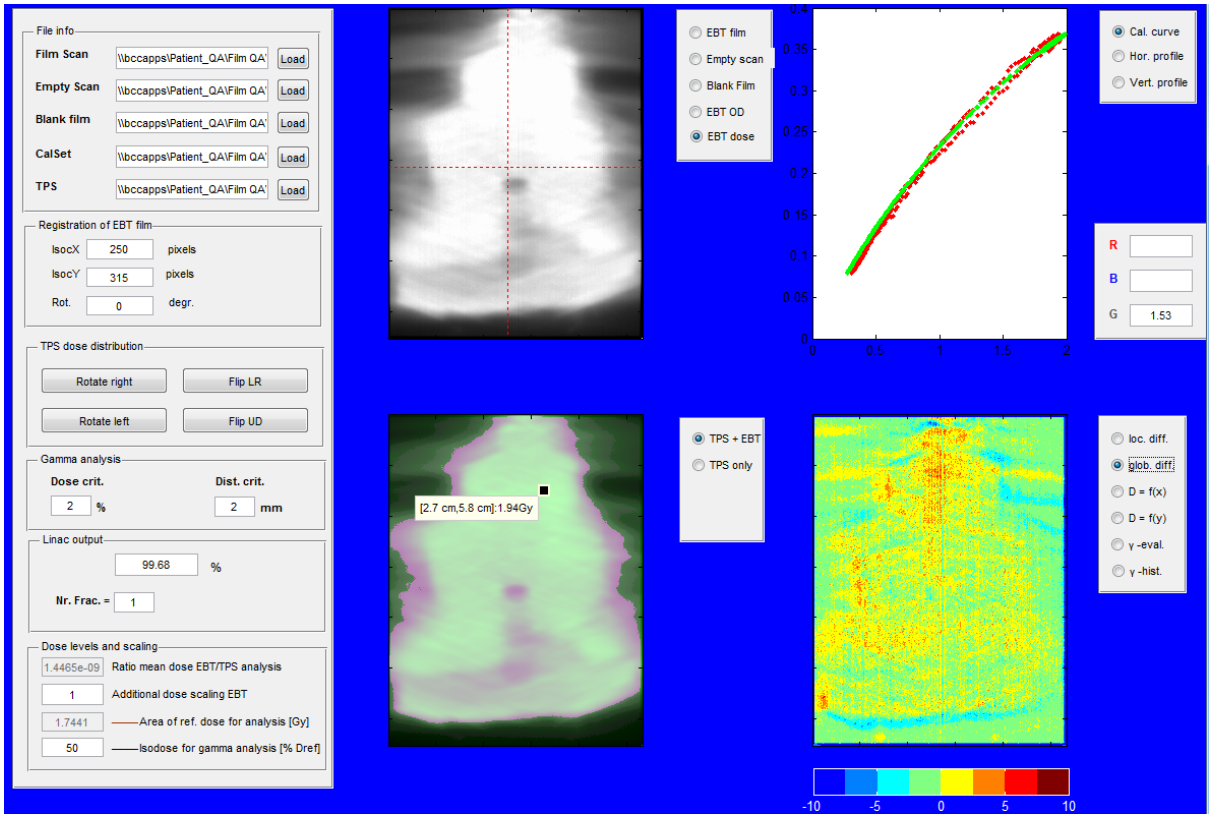

Figure 2.13: *Screenshot of the EBT film analysis software application. Each panel has multiple functions that can be selected by the user. In this example, the various panels show the measured film dose (top left panel), the calibration curve used to convert from film OD to dose (top right panel), an overlay of the film dose on the TPS dose (bottom left panel), and the global difference plot of the film dose minus the TPS dose (bottom right panel).*

2.6.4. ArcCheck Measurements

Equipment

QC measurements were performed using an ArcCheck cylindrical diode array phantom [75, 76], which consists of 1386 diodes arranged helically at a depth of 2.9 cm in a cylindrical acrylic (PMMA) phantom. Adjacent diodes are positioned 1.0 cm from one another. An acrylic insert was used to fill the air gap in the centre of the cylindrical array. The SNC Patient software (Sun Nuclear Corp., Melbourne FL, USA) was used to control the device during measurement and also for data analysis and comparison with TPS calculated dose.

TPS Dose Calculation – Using Standard WBCC Set Up

The TPS dose was calculated by producing a verification plan from the clinical patient plan. This was done by copying the clinical VMAT plan onto a CT data set of the ArcCheck phantom, and recalculating the plan dose for this configuration using both the clinical and adjusted beam models. No Hounsfield-Unit override (HUo) was used for the ArcCheck phantom, and the ArcCheck heterogeneity correction (HC) was turned off in the SNC Patient software for the standard WBCC ArcCheck set up. A virtual couch structure was used in Eclipse to simulate the attenuation through the couch with the couch surface given a value of -450 HU and the body of the couch having a value of -970 HU at the WBCC. The centre of the ArcCheck phantom was aligned to the treatment machine isocentre. The TPS dose was then calculated using a 1.5 mm dose grid resolution. The resultant VMAT plan and dose were exported from Eclipse as DICOM RT files and imported into the SNC patient software to be used for analysis.

TPS Dose Calculation – Using Manufacturer Recommended Set Up

As mentioned in the above paragraph, the standard WBCC ArcCheck set up does not utilise an HUo or the ArcCheck HC. The manufacturer does recommend using these corrections to provide a better agreement between measured and calculated dose distributions by reducing the effect of streaking artefacts for the high density diodes in the CT scan of the ArcCheck leading to elevated HU values that can then influence the dose calculation [76]. Furthermore, the HU value for PMMA does not get correctly converted to electron density using a clinical CT to electron density calibration curve. Therefore ArcCheck dose calculations were repeated with these corrections implemented.

The ideal HU override for the ArcCheck phantom has previously been determined at the WBCC and has a value of 280 HU [77] (see **Figure 2.14**). To apply the HC, the SNC patient software accesses a file containing predetermined heterogeneity correction factors for all diodes while it is post-processing a completed measurement. Therefore, the HC cannot be retrospectively applied and this correction has to be turned on before collecting data. This required recollecting all measurement data

again with the HC turned on. All recollected data was then compared to the error-free verification plans calculated using the HU override for the ArcCheck phantom.

Before the ArcCheck plans could be re-measured with the HC applied, the DLG of the linac used for measurements was altered outside this study for clinical purposes. Therefore, all ArcCheck verification plans with the HUo applied were calculated using a different beam model. For this beam model, the DLG was set to 1.6 mm and the ETSS was set to 1.5 mm in the X direction and 0.0 mm in the Y direction.

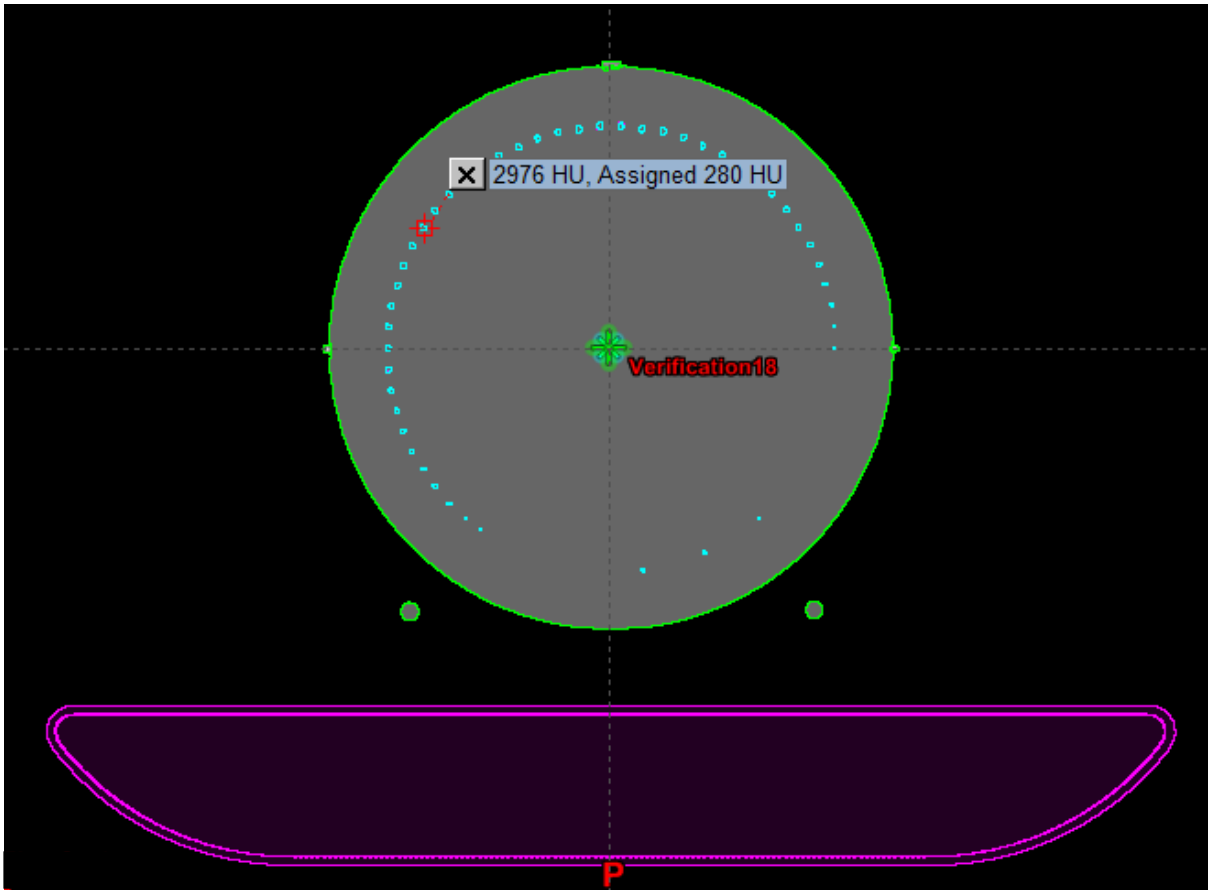

Figure 2.14: *CT scan of the ArcCheck phantom with the Hounsfield unit override applied The HU value for the selected diode in the red box was determined to be 2976 HU according to the CT scan, but has been overridden to a value of 280.*

ArcCheck Calibration

The ArcCheck absolute dose calibration was performed by irradiating the two calibration diodes on the top of the ArcCheck array with a 6 MV 10 x 10 cm^2 field to 200 MU. The AAA TPS dose calibration of the reference set up was used to determine that 267cGy.200MU^{-1} (264 cGy.200MU^{-1} with the HUo and HC applied) was delivered to the calibration diodes. This value was entered into the SNC patient software during the absolute dose calibration procedure. During the absolute calibration procedure the output of the linac was measured using a Farmer ionisation chamber placed on the central axis of the ArcCheck device to ensure the linac output was within tolerance. The ionisation chamber was removed prior to delivery of any VMAT plans. The relative responses of the remaining diodes were defined using an annual array calibration procedure. Both the absolute dose calibration and the relative array calibration were verified immediately at the start of each measurement session with the ArcCheck.

ArcCheck Plan Measurement

The ArcCheck device was positioned on the thick section of the IGRT exact treatment couch, and the positioning lines on the exterior of the ArcCheck phantom were aligned to the in room lasers and linac cross hair. Once positioned and calibrated, the dose delivered to the ArcCheck using the verification plan was measured for all error-free plans and plans containing introduced errors, and the resultant helical dose map was saved for comparison with the TPS dose map using the SNC Patient software.

Analysis

The TPS DICOM RT dose file was imported into the SNC Patient software and was converted by the software into a helical dose map to match the geometry of the ArcCheck diodes. Agreement between measured and calculated dose was carried out using global γ analysis with a low dose threshold of 10% [74] and the results were evaluated using two sets of acceptance criteria. The current WBCC acceptance criteria for ArcCheck QC are:

i. $\{2\%, 2\ mm\}\ \gamma$ passing rate $> 85\%$
ii. $\{3\%, 3\ mm\}\ \gamma$ passing rate $> 95\%$

2.7. Sensitivity Analysis

Three separate sensitivity metrics were used throughout the course of this study. They are designated S_1, S_2, and S_3 respectively and are defined in detail below:

2.7.1. S_1

Commonly, sensitivity has been defined in a statistical sense based on dichotomous classifications i.e. positive or negative test outcome [78]. Then, based on an assigned 'true state' of the outcome i.e. whether a particular condition is present or not, four outcomes are possible;

1) True positive - TP (positive test result when the condition is present)
2) True negative - TN (negative result when the condition is not present)
3) False positive - FP (positive result when the condition is not present)
4) False negative - FN (negative result when the condition is present)

From these results, sensitivity is defined as the true positive rate (TPR) and is calculated using Equation 2.8:

$$S_1 = TPR = \frac{TP}{TP+FN} \qquad \text{Equation 2.8}$$

S_1 is used commonly in radiotherapy to determine the sensitivity of QC tests to plans where errors have been intentionally introduced [45 - 47, 50]. Therefore, this metric provides the best ability to compare the QC methods at the WBCC with other studies. It is important to consider that by defining sensitivity in this way, the QC result depends on the measurement method as well as the user defined passing criteria.

Both the film and ArcCheck QC methods contain two separate metrics to determine if a plan passes the QC. For film these are the γ pass rate for the {2%;2mm} criterion must be above 85% and the

mean value of γ must be less than 0.5. For ArcCheck, the passing criteria are 95% of points must pass the {3%;3mm} γ-criterion and 85% of points must pass the {2%;2mm} criterion. A conservative approach was applied in this study for both these QC methods by labelling a failure for either criterion as a positive result, while only classifying a result as negative when both acceptance criteria are met.

In addition to the definition of sensitivity based on dichotomous classification of test results, sensitivity of the QC measurement methods can also be defined as the ratio of the change in output from the test over the change in input [42]. For the purposes of this study with regard to patient-specific QC, the change in output is defined as the change in QC result for a plan containing an introduced error compared to the treatment plan without any errors. The change in input can be defined in different ways and results in metrics S_2 and S_3.

2.7.2. S_2

S_2 is a metric that was developed to compare the sensitivity of the QC method for different error modes. The change in input is defined as the difference between the TPS calculated verification plan of the plan containing the intentional error and the TPS calculated verification plan of the plan without any introduced errors. For the point dose QC method, S_2 was calculated using:

$$S_2 = \frac{Dose_{meas.}^{error} - Dose_{meas.}^{orig}}{Dose_{TPS}^{error} - Dose_{TPS}^{orig}} \qquad \text{Equation 2.9}$$

Where:

- $Dose_{meas.}^{error}$ is the dose measured using the trPD method for a plan containing an error.
- $Dose_{meas.}^{orig}$ is the dose measured using the trPD method for the corresponding error-free plan.
- $Dose_{TPS}^{error}$ is the dose calculated by the TPS in the slab phantom at the same point as was measured for the verification plan containing the error.
- $Dose_{TPS}^{orig}$ is the dose calculated by the TPS in the slab phantom at the same point as was measured for the corresponding error-free verification plan.

By definition S_2 is zero for intentional TPS errors, therefore S_2 was determined only for delivery errors.

For this metric the change in output was made to be the same quantity as the change in input i.e. for the point dose QC method, the change in output (numerator) will be the change in dose at a particular point in the phantom. The change in input (denominator) was then selected to be the change in dose calculated at the corresponding point in the verification plan by the TPS. This enabled comparison of the sensitivity between different error modes for a given QC method.

For analysis of film and ArcCheck results, the calculation of S_2 was based on a dose difference analysis using the γ analysis function of the Film or ArcCheck software while applying either a 2% or a 3% dose difference criterion and a DTA criterion of 0 mm. This effectively converts the γ analysis into a dose difference analysis. For the film and ArcCheck QC methods S_2 was defined using Equation 2.10:

$$S_2 = \frac{P^{\{X\%\,DD\}}_{error,\ meas.} - P^{\{X\%\,DD\}}_{orig,\ meas.}}{P^{\{X\%\,DD\}}_{error,\ calc.} - P^{\{X\%\,DD\}}_{orig,\ calc.}} \qquad \text{Equation 2.10}$$

Where:

- $P^{\{X\%\,DD\}}_{error,\ meas.}$ is the percentage of points on the measurement dose map passing an X% dose difference criterion for a plan containing an error.
- $P^{\{X\%\,DD\}}_{orig,\ meas.}$ is the percentage of points on the measurement dose map passing an X% dose difference criterion using the film method for the corresponding error-free plan.
- $P^{\{X\%\,DD\}}_{error,\ calc.}$ is the percentage of points in the corresponding dose map (calculated by the TPS for film or SNC Patient for the ArcCheck) that pass an X% dose difference criterion for the verification plan containing the error.
- $P^{\{X\%\,DD\}}_{orig,\ calc.}$ is the percentage of points in the corresponding dose map (calculated by the TPS for film or SNC Patient for the ArcCheck) that pass an X% dose difference criterion for the corresponding error-free verification plan.

2.7.3. S_3

The final sensitivity metric used for this study was related to the specific error introduced to the plan. In this case the change in input was the magnitude of the introduced error. S_3 was defined as the ratio of change in QC result over the magnitude of the introduced error and calculated using Equation 2.11:

$$S_3 = \frac{\Delta QC\ Result}{m} \qquad \text{Equation 2.11}$$

Where:

- *m* is the magnitude of the introduced error.

S_3 is expressed either in %.mm^{-1} (for MLC shift errors) or is dimensionless (for MU errors). Using this method it was possible to determine how much the QC results change per a certain change in error magnitude, and to determine at what magnitude an error becomes detectable.

2.8. Specificity Analysis

Two specificity metrics were used throughout the course of this study. They were designated Sp_1 and Sp_2 respectively and are defined in detail below:

2.8.1. Sp_1

Sp_1 is based on the same methodology as S_1 above (See section 2.7.1). In this case, specificity is defined as the true negative rate (TNR) and is calculated using Equation 2.12:

$$Sp_1 = TNR = \frac{TN}{TN+FP} \qquad \text{Equation 2.12}$$

This definition of specificity represents the ability of the QC method to pass a QC measurement of the treatment plan containing no introduced errors.

2.8.2. Sp_2

One of the main areas of interest in this study was to determine if the QC methods could be used to resolve what type of error has occurred. Since each QC method used a different analysis, separate Sp_2 methods would be required for each QC method. Due to time constraints, a method for Sp_2 analysis was only developed for trPD measurements.

For trPD measurements, this could be achieved by utilising the dose per CP data and plotting the difference between measured and calculated segment dose against the detector DTFE (see section 2.6.1). Using this analysis method average deviation of all points in each of the three regions (see **Figure 2.10**) can then be calculated.

Since MLC positional errors affect the field size, the majority of deviations caused by these errors would likely occur near the field edge i.e. in region II, whereas MU errors are likely to have a larger effect when the detector is in the open field i.e. in region III. Therefore, by studying the average dose deviation in each region, the ability of the trPD method to resolve whether or not a detected error was caused by a MLC shift (or potentially a DLG shift), or by a change in output/change in MU delivered (either systematic, or varied with gantry angle) was investigated.

2.9. Receiver Operator Characteristic (ROC) Curve Analysis

Receiver operator characteristic (ROC) curves are commonly used to determine the optimal sensitivity (S_1) and specificity (Sp_1) of a diagnostic test method and to evaluate the efficiency of the test method [79 - 81]. These curves were generated by calculating S_1 and Sp_1 while varying the QC passing criterion over a wide range. The resulting S_1 and Sp_1 values were then plotted yielding an ROC curve as exemplified in **Figure 2.15**. A non-informative test method results in an ROC curve

equivalent to a diagonal line, while a perfect test method would result in an ROC curve displaying both 100% sensitivity and 100% specificity for a given configuration (top left hand corner of **Figure 2.15**). The area under the curve (AUC) is a common metric for determining the overall efficiency of a diagnostic test method. An AUC of 0.5 corresponds to a non-informative test method while a value of 1.0 corresponds to a perfect test method. The ROC curve can also be used to determine the QC acceptance criterion or 'cut-off value' for the test which provides the optimal S_1 and Sp_1. This was conducted by using the Youden index (J), which defines the point on the ROC curve at the maximum distance from the diagonal line representing AUC = 0.5 diagonal line. The Youden index is calculated using Equation 2.13:

$$J = \max\left[S_1(c) + Sp_1(c) - 1\right] \qquad \text{Equation 2.13}$$

Where:

- c is a given cut-off value on the ROC curve [82].

For this study, ROC analysis including AUC calculation and determination of optimal cut-off criteria was conducted using software developed in house using Matlab (see appendix 2.B). This software supported the ROC analysis for either a single error mode and magnitude, or a range of error modes and magnitudes. Using this ROC analysis methodology, S_1 and Sp_1 were able to be compared using four separate configurations:

A. Clinically applied TPS beam model and QC acceptance criteria.

B. Adjusted TPS beam model and currently applied QC acceptance criteria.

C. Clinically applied TPS beam model and optimised QC acceptance criteria.

D. Adjusted TPS beam model and optimised QC acceptance criteria.

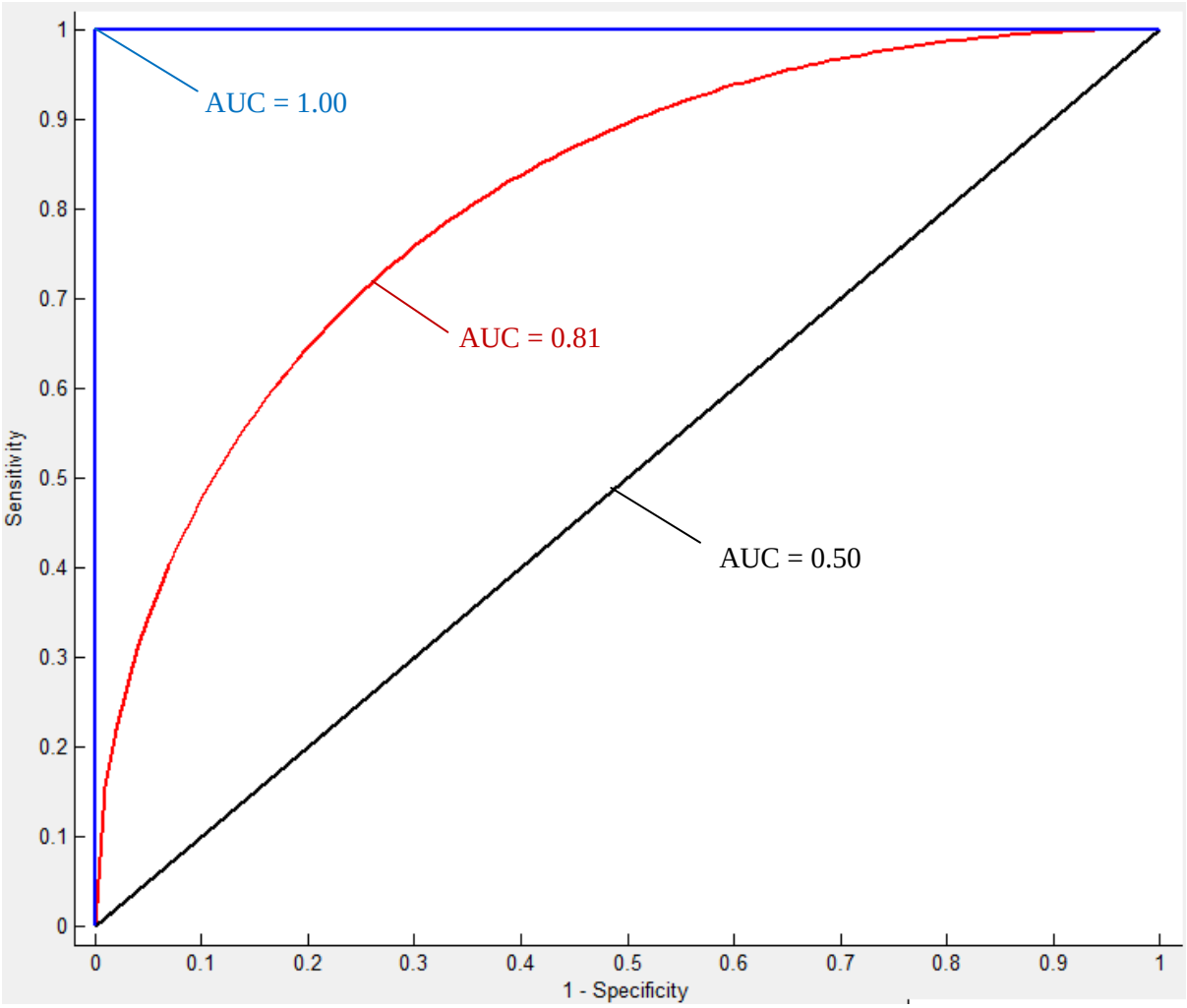

Figure 2.15: *Example ROC curves. The black line represents a non-informative test method, the blue line represents the perfect test method, and the red line represents a test method with an AUC between 0.5 and 1.0. Note that the specificity scale (x axis) is 1 – specificity and runs from 0.0 to 1.0.*

3. Results

The results obtained throughout this study are presented in this chapter in the following sections:

- Section 3.1 presents the results of the introduced error planning study are detailed in terms of their effect on the DVH metrics.
- Sections 3.2 to 3.5 present the results for each QC method are presented including:
  - The QC measurement results for error-free plans.
  - The QC measurement results for plans containing introduced errors.
  - The sensitivity and specificity analysis using all metrics defined in sections 2.7 and 2.8.

Sections 3.2 to 3.5 refer to verification plans calculated using two separate beam models, the 'clinical beam model' and the 'adjusted beam model' as discussed at the beginning of chapter 2 (for more detail about the beam model adjustment see appendix 2.A for methodology and 3.A for results). The only parameters that were different between these two models were the DLG (2.0 mm in the clinical beam model, 1.2 mm in the adjusted model) and the ETSS in the X direction (0.0 mm in clinical beam model, 1.5 mm in the adjusted beam model). All other parameters and beam data were the same for both beam models.

The metrics S_1 and Sp_1 were determined for each QC method using the 4 different configurations (denoted **A - D**) defined in section 2.9. In addition, a comparison was made between the current setup of the ArcCheck system and the vendor recommended setup including HU override and heterogeneity corrections. These two ArcCheck-configurations will be labelled **A'** and **C'** as they are modifications of ArcCheck configurations **A** and **C**.

3.1. Assessing Clinical Relevance of Introduced Errors

All DVH metrics in section 3.1 are based on calculations using the clinical beam model as discussed in section 2.2. The clinical relevance of each error was determined using the criteria detailed in

section 2.2 and was assumed to be the same regardless of which beam model was used for calculations. A summary of the TPS calculated DVH metrics for all introduced errors are given in **Table 3.1**.

Table 3.1: *Change in DVH metrics for all introduced errors. Each pair of rows displays the median and range of observed DVH changes over N=5 patients for a specific error magnitude. Orange highlighting indicates for which ROIs the clinical relevance criteria were violated.*

Error	Metric	PTV66		SC	BS	Chiasm	No. of plans where clinically relevant
		ΔD_{98} (%)	ΔD_{1cc} (%)	ΔD_2 (%)	ΔD_2 (%)	ΔD_2 (%)	
3% decrease	Median	**-2.9**	-3.2	-2.0	-2.2	-2.0	**5**
	Range	**(-2.9 ; -2.9)**	(-3.2 ; -3.1)	(-2.0 ; -1.8)	(-2.8 ; -1.6)	(-2.7 ; -0.2)	
1.5 % decrease	Median	-1.4	-1.6	-1.0	-1.1	-1.0	**0**
	Range	(-1.4 ; -1.4)	(-1.6 ; -1.6)	(-1.0 ; -0.9)	(-1.4 ; -0.8)	(-1.4 ; -0.1)	
1.5 % increase	Median	1.4	1.6	1.0	1.1	1.0	**0**
	Range	(1.4 ; 1.5)	(1.6 ; 1.6)	(0.9 ; 1.0)	(0.8 ; 1.4)	(0.1 ; 1.4)	
3 % increase	Median	2.9	**3.2**	**2.0**	**2.2**	**2.0**	**5**
	Range	(2.9 ; 2.9)	**(3.1 ; 3.2)**	(1.8 **; 2.0)**	(1.6 **; 2.8)**	(0.2 **; 2.7)**	
1.0 mm Closed Shift	Median	**-4.6**	-3.0	-3.4	-3.9	-7.1	**5**
	Range	**(-6.3 ; -3.7)**	(-4.8 ; -2.3)	(-3.7 ; -2.5)	(-7.3 ; -2.8)	(-7.2 ; -0.3)	
0.5 mm Closed Shift	Median	**-2.1**	-1.5	-1.6	-2.0	-3.3	**3**
	Range	**(-2.9 ;** -1.7)	(-2.5 ; -1.2)	(-1.9 ; 0.4)	(-3.5 ; -1.4)	(-3.7 ; -0.1)	
0.5 mm Open Shift	Median	2.2	**2.2**	**2.1**	**2.3**	**3.9**	**3**
	Range	(1.6 ; 2.8)	(1.5 **; 3.1)**	(1.6 **; 2.3)**	(1.7 **; 3.8)**	(0.2 **; 4.3)**	
1.0 mm Open Shift	Median	4.1	**4.6**	**3.9**	**4.7**	**7.7**	**5**
	Range	(2.9 ; 5.2)	**(2.9 ; 6.5)**	**(2.9 ; 4.7)**	**(3.2 ; 7.0)**	(0.4 **; 8.2)**	
1.0 mm Translation	Median	-0.5	0.3	0.9	0.6	0.0	**0**
	Range	(-0.9 ; -0.5)	(0.1 ; 0.5)	(0.1 ; 1.0)	(-1.0 ; 1.1)	(-1.0 **; 2.4)**	
2.0 mm Open near BS	Median	0.8	**4.5**	0.5	**6.8**	**5.6**	**5**
	Range	(0.7 ; 3.0)	**(3.6 ; 5.9)**	(0.4 ; 1.6)	**(5.3 ; 10.6)**	(0.4 **; 8.5)**	
2.0 mm Open near Chiasm	Median	0.1	0.8	0.0	0.0	**7.2**	**2**
	Range	(0.0 ; 0.2)	(0.1 - 1.8)	(-0.1 ; 0.0)	(-0.1 ; 0.3)	**(4.7 ; 9.9)**	
2.0 mm Open near SC	Median	1.0	**2.8**	**6.0**	1.0	0.0	**5**
	Range	(0.7 ; 1.9)	(1.8 **; -4.6)**	**(4.5 ; 8.1)**	(0.2 ; 1.8)	(0.0 ; 0.1)	
1.0 mm Open near SC	Median	0.7	1.2	**2.9**	0.6	0.2	**3**
	Range	(0.0 ; 1.1)	(0.7 ; 1.4)	(2.3 ; 3.3)	(0.2 ; 1.0)	(0.0 ; 0.2)	
4% decrease at gantry 180°	Median	-1.2	-1.2	-0.6	-1.0	-0.4	**0**
	Range	(-1.2 ; -1.0)	(-1.3 ; -0.9)	(-0.9 ; -0.5)	(-1.7 ; -0.5)	(-0.7 ; -0.1)	
8% decrease at gantry 180°	Median	**-2.4**	-2.3	-1.2	-2.0	-1.0	**5**
	Range	**(-2.7 ; -2.1)**	(-2.8 ; -1.8)	(-1.9 ; -1.0)	(-3.2 ; -1.2)	(-1.6 ; -0.1)	
ETSS X set to 1.5 mm	Median	0.0	0.0	0.0	0.0	0.0	**0**
	Range	(0.0 ; 0.0)	(0.0 ; 0.0)	(0.0 ; 0.1)	(-0.3 ; 0.0)	(0.0 ; 0.1)	
DLG set to 1.2 mm	Median	-1.7	-1.3	-1.4	-1.8	-2.9	**1**
	Range	**(-2.5 ;** -1.4)	(-2.2 ; -1.1)	(-1.6 ; -1.1)	(-3.0 ; -1.1)	(-3.4 ; -0.2)	

For all patients, introduced MU errors of ±3% violated the clinical relevance criterion for at least 1 DVH metric. Consequently, all these errors were clinically relevant for all patients. MU errors of ±1.5% did not violate any clinical relevance criteria for any DVH metric. The MU errors of this magnitude were therefore not clinically relevant for any of the patients in this study.

For the 8% machine output errors that varied with gantry angle, the change in PTV D_{98} was greater than 2% for all patients. Therefore this error was clinically relevant for all patients. The 4% output error that varied with gantry angle did not result in a decrease in D_{98} of more than 2% for any patient, and was not considered clinically relevant for any patient in this study.

For systematic shifts of the entire leaf banks:

- Both 1.0 mm open and closed shifts did violate the clinical relevance criterion for at least 1 DVH metric, and were therefore clinically relevant for all patients.
- The 0.5 mm open and closed shifts violated the clinical relevance criterion for at least 1 DVH metric for three out of five patients. Therefore the clinical relevance of these errors was plan dependent. The 1.0 mm translation shift did not breach any DVH criterion for any patient and was not clinically relevant for any patient in this study.

For MLC shifts near OARs:

- The 2.0 mm shift near the SC violated the clinical relevance criteria for at least one metric for all 5 patients. For four patients, the D_2 dose constraint for the SC ($D_2 < 45$ Gy) was exceeded, while for patient 3, this error was clinically relevant because the PTV change in D_{1cc} was more than 2%. Therefore this error was clinically relevant for all 5 patients
- The 2.0 mm shift near the BS violated the clinical relevance criteria for at least one metric for all 5 patients. For four patients, the BS D_2 dose constraint ($D_2 < 50$ Gy) was exceeded, while

for patient 3, this was clinically relevant because the PTV change in D_{1cc} was more than 2%. Therefore this error was clinically relevant for all 5 patients

- The 2.0 mm shift near the chiasm caused a breach of clinical relevance criteria for one DVH metric for 2 out of the 4 patients this error was introduced for. Therefore it was clinically relevant for 2 out of 4 patients.
- The 1.0 mm shift near the SC led to a violation of clinical relevance criteria for 3 out of 5 patients. Therefore this error was clinically relevant for 3 out of 5 patients.

The most prominent effect of decreasing the DLG from 2.0 mm to 1.2 mm was an overall reduction of the calculated dose. For the PTV ΔD_{98}, it resulted in a median reduction of -1.7% (range -1.4 to -2.5%). There was one individual patient where the ΔD_{98} was more than 2% (Patient 3, ΔD_{98} = -2.5%), so this error was clinically relevant for 1 out of 5 patients. This highlighted that reducing the DLG by 0.8 mm may cause a clinically relevant error but it is patient dependent. The overall change in calculated dose due to the change in DLG also resulted in the same difference in calculated dose between configurations **A** and **B**.

It was found that changing the size of the effective target spot size in the x direction by 1.5 mm was not clinically relevant for any of the five patients in this study.

3.2. trPD Results

Error-free Plan Verification

Each patient plan without introduced errors was measured using the trPD technique on three separate occasions and included a number of repeat measurements to verify the reproducibility of the results. Overall, the PTV point was measured three times, the SC PRV point twice, and the BS PRV and

chiasm points were measured once. The QC result and indication of whether each measurement was considered a true negative or false positive is given in **Table 3.2**.

The measured integral dose at the PTV reference point and the dose calculated by the TPS agreed within ±2.0%. There was also very good reproducibility of PTV measurements with the maximum difference across measurement sessions for the same measurement of 0.7% (Patients 3 and 5).

The agreement between measured and TPS dose for OAR measurement points was poor in contrast to the PTV measurements. 60% of SC PRV measurements, 80% of BS PRV measurements and 100% of chiasm measurements did not agree with the TPS calculated dose within ±2.0%, so were therefore classified as false negatives. This is consistent with previous results in our department at the OAR measurement locations that were situated in high dose gradients. The maximum dose gradient at the OAR locations was $8\%.mm^{-1}$. With an estimated positioning uncertainty of 0.5 - 1.0 mm, the estimated uncertainty in dose was 4 - 8%. Therefore, the finite positioning accuracy of the detector largely explained the poorer results for the OAR locations. In addition, volume averaging in the longitudinal direction of the detector may play a role as well [51] . The results for the PTV and OAR points will be analysed separately in the remainder of this study to highlight the difference between the results for these locations.

Table 3.2: *Integral point dose difference (Δ in %) for measurement of the error-free plans (see Equation 2.3) using configuration* ***A****. False Positive results are bolded and highlighted in orange; all other results are TNs.*

Patient	ROI	Measurement 1	Measurement 2	Measurement 3
		Δ [%]	Δ [%]	Δ [%]
Patient 1	PTV	-0.6%	-0.6%	-0.4%
	SC PRV	-1.4%	**-3.0%**	-
	BS PRV	**-3.7%**	-	-
	Chiasm	-	-	-
Patient 2	PTV	-0.2%	-0.3%	0.1%
	SC PRV	**-2.3%**	**-3.1%**	-
	BS PRV	**-3.1%**	-	-
	Chiasm	**-14.4%**	-	-
Patient 3	PTV	0.8%	0.8%	1.4%
	SC PRV	**-2.1%**	-1.4%	-
	BS PRV	**-5.1%**	-	-
	Chiasm	**-3.9%**	-	-
Patient 4	PTV	-1.6%	-1.8%	-1.7%
	SC PRV	**-3.9%**	**-2.1%**	-
	BS PRV	**-9.6%**	-	-
	Chiasm	**-14.9%**	-	-
Patient 5	PTV	-0.8%	-0.9%	-0.2%
	SC PRV	-1.3%	-0.6%	-
	BS PRV	-0.7%	-	-
	Chiasm	**-25.0%**	-	-

Verification of Plans Including Intentional Errors

The measured integral dose deviation and dichotomous classification for all results are summarised in **Table 3.3**.

Table 3.3: *QC results and indication of outcome for all plans containing introduced errors for configuration **A**. Orange shading corresponds to false negatives and purple shading corresponds to false positives. All non-shaded results are either true negatives or true positives.*

Plan error(s)	Patient 1		Patient 2		Patient 3		Patient 4		Patient 5	
	Δ [%]	Outcome	Δ [%]	Outcome	Δ [%]	Outcome	Δ [%]	Outcome	Δ [%]	Outcome
MU 3 % low	-3.3%	TP	-2.9%	TP	**-1.5%**	**FN**	-4.9%	TP	-3.2%	TP
MU 1.5% low	-1.8%	TN	-1.6%	TN	0.0%	TN	**-3.4%**	**FP**	-1.7%	TN
MU 1.5% high	1.0%	TN	1.4%	TN	**2.5%**	**FP**	-0.8%	TN	0.6%	TN
MU 3% high	2.4%	TP	2.8%	TP	4.1%	TP	**1.3%**	**FN**	2.2%	TP
Output w gantry angle 8%	-2.4%	TP	-2.5%	TP	**-0.5%**	**FN**	-4.3%	TP	-2.9%	TP
Output w gantry angle 4%	-1.6%	TN	-1.4%	TN	-0.1%	TN	**-3.5%**	**FP**	**-2.2%**	**FP**
MLC closed 1 mm	-2.3%	TP	-3.6%	TP	-3.8%	TP	-5.9%	TP	-4.3%	TP
MLC closed 0.5 mm	-1.8%	TN	-1.9%	TN	**-1.2%**	**FN**	-4.4%	TP	-2.7%	TP
MLC open 0.5 mm	0.7%	TN	**1.7%**	**FN**	3.3%	TP	**-0.1%**	**FN**	1.0%	TN
MLC open 1 mm	2.0%	TP	3.4%	TP	5.6%	TP	2.7%	TP	3.0%	TP
MLC translation 1mm	-0.9%	TN	-1.1%	TN	1.6%	TN	**-2.6%**	**FP**	-1.5%	TN
MLC SC 1 mm - PTV	**-0.4%**	**FN**	**0.4%**	**FN**	1.1%	TN	**-1.7%**	**FN**	-0.3%	TN
MLC SC 1 mm - SC	**0.6%**	**FN**	**-0.2%**	**FN**	**2.2%**	**FP**	**1.9%**	**FN**	**2.6%**	**FP**
MLC SC 2 mm - PTV	**-0.3%**	**FN**	**1.0%**	**FN**	**1.2%**	**FN**	**-1.2%**	**FN**	**0.7%**	**FN**
MLC SC 2mm - SC	6.2%	TP	4.0%	TP	5.5%	TP	4.6%	TP	5.2%	TP
MLC BS 2 mm - PTV	2.4%	TP	5.8%	TP	5.7%	TP	2.9%	TP	3.9%	TP
MLC BS 2 mm - BS	4.4%	TP	7.4%	TP	5.9%	TP	7.0%	TP	4.1%	TP
MLC Chiasm 2 mm - PTV	-		**-0.2%**	**FN**	1.1%	TN	**-1.6%**	**FN**	-0.7%	TN
MLC Chiasm 2 mm - Chiasm	-		-4.5%	TP	**3.9%**	**FP**	-3.3%	TP	-1.8%	TN
MU 3% high, MLC 1mm closed	-0.3%	TN	-0.8%	TN	**1.0%**	**FN**	-3.4%	TP	-1.6%	TN
DLG 1.2 mm	0.4%	TN	1.3%	TN	2.8%	TP	-0.3%	TN	0.8%	TN
ETSS X 1.5 mm	-0.6%	TN	-0.2%	TN	1.0%	TN	-2.0%	TN	-0.8%	TN

These results are summarised in truth tables (including the 34 measurements of the plans without errors) for configurations **A** and **B** in **Table 3.4A - B** for PTV measurement points and **Table 3.5A - B** for OAR measurement locations.

Table 3.4A-B: *Truth tables based on the clinical relevance of introduced errors as defined in section 2.2 for the trPD measurements at the PTV measurement points.*

Configuration **A**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	32	17
Not clinically relevant	5	50

Configuration **B**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	31	18
Not clinically relevant	11	44

Table 3.5A-B: *Truth tables based on the clinical relevance of introduced errors as defined in section 2.2 for the trPD measurements at the OAR measurement points.*

Configuration **A**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	12	3
Not clinically relevant	17	6

Configuration **B**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	12	3
Not clinically relevant	8	15

For configuration **A** the percentage of FPs and FNs was considerably higher for the OAR locations (53%) compared to the PTV locations (21%), while it was similar for configuration **B** (28% at PTV locations, 29% at OAR locations). The majority of the false positives were due to the difficulty in measuring dose at the OAR locations; all the selected OAR locations had much larger dose gradients than the PTV locations due to the fact that the close proximity of PTV and OAR necessitated a rapid fall off in dose from the PTV boundary towards the OAR in order to meet the planning constraints so any set up error could have a large effect on the measurement. This could account for a number of the false negative results if an introduced error increased the dose at the OAR measurement point such that it went from being more than 2% below the TPS dose to within 2% of the TPS dose. Furthermore for OAR specific MLC errors, the PTV measurement point is not close to the location which was effected by the error, which could account for a number of false negatives.

3.2.1. S_1 and Sp_1

Figure 3.1A - B graphically illustrate for the PTV points the relationship between QC results and clinical relevance, how the corresponding acceptance criteria define whether results are either positive or negative and either true or false for two DVH metrics. The clinical relevance criteria and the QC acceptance criteria define the 4 separate regions corresponding to TPs, TNs, FPs or FNs. **Figure 3.1A - B** shows that the majority of points are located in the TP and TN regions for the results measured at the PTV location (83 true results and 23 false results). The linear regression line of each plot shows that the trPD QC results have a moderate correlation with the chosen clinical relevance metrics (R^2 = 0.65 to 0.71).

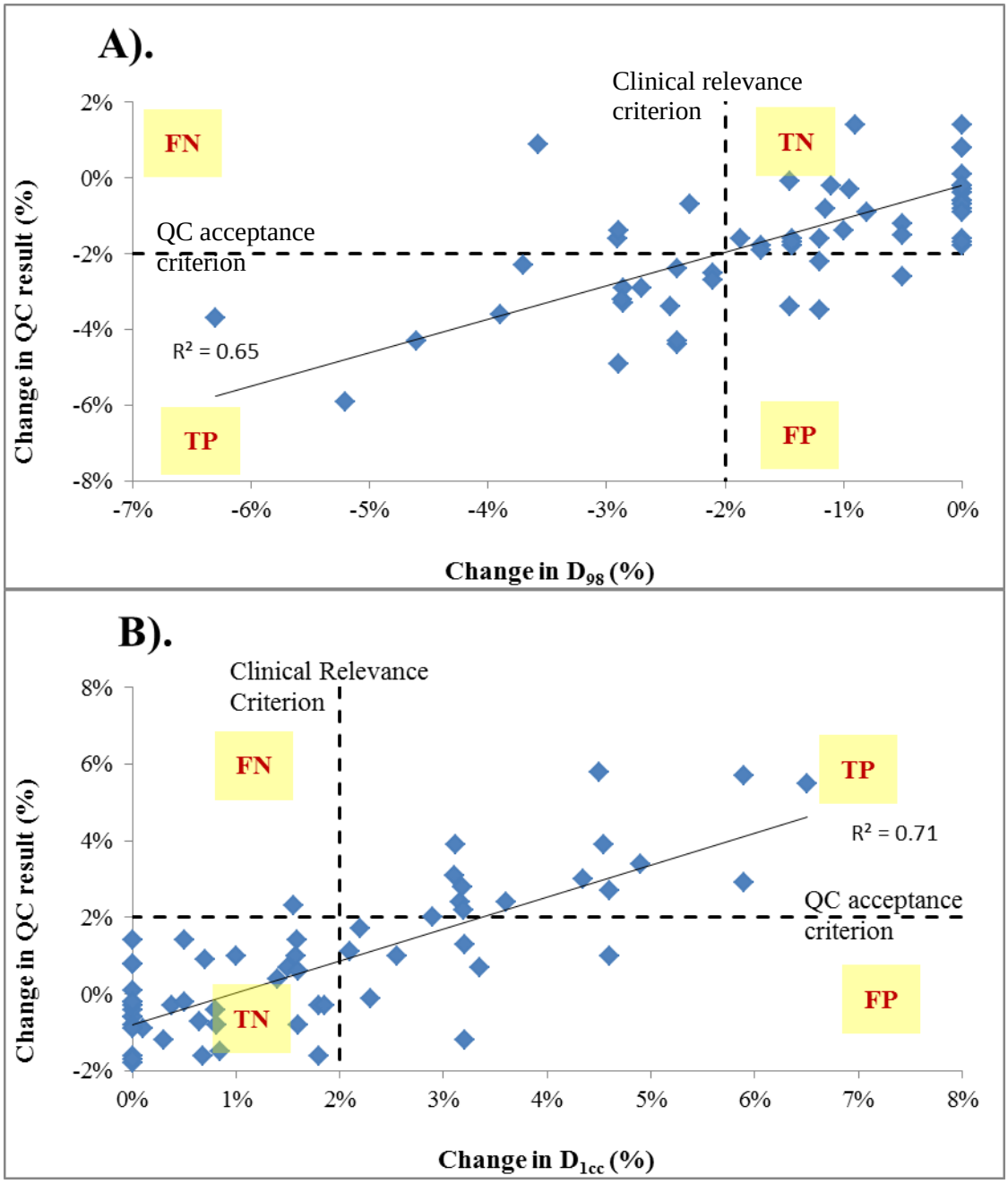

Figure 3.1 A (top) and B (bottom): *QC results based on TPS calculations using configuration **A** plotted as a function of the change in a DVH metric for PTV point verification measurements of plans including intentional errors. The horizontal and vertical dotted lines indicate the QC passing criterion and clinical relevance criterion, respectively. These criteria define the regions of true and false positives, and true and false negatives.*

Subsequently, S_1 and Sp_1 metrics were calculated for PTV and OAR point measurements after including all DVH metrics defining clinical relevance (**Table 3.1**) using Equation 2.8 and Equation 2.12 and these values are displayed in **Table 3.6**.

Table 3.6: *S_1 and Sp_1 for the point dose method as determined using the methods outlined in sections 2.7.1 and 2.8.1 for both configuration **A** and **B** for both PTV measurement points and OAR measurement points.*

PTV	Configuration A	Configuration B
S_1	65.3%	63.3%
Sp_1	90.9%	80.0%

OAR	Configuration A	Configuration B
S_1	80.0%	80.0%
Sp_1	26.1%	65.2%

The obtained value of Sp_1 for PTV locations was high, as was the value of S_1 at OAR locations. However, Sp_1 for OAR measurement locations showed that only one quarter (Configuration A) to two thirds (Configuration B) of the error-free plans pass the QC. This low specificity is likely to be due to the high dose gradients present at most OAR locations. This subject is discussed in more detail in sections 4.2 and 4.3.

In contrast to the high Sp_1 values for the PTV locations, the obtained S_1 values showed that only two thirds of errors were detected regardless of the applied configuration, which was lower than anticipated. There are two factors that could potentially cause this difference:

1) The applied QC acceptance criteria may be sub-optimal.
2) Inclusion of PTV point verification measurements of intentional error plans that were specifically aiming to influence the dose delivery to an OAR.

Both factors were further investigated as described below.

The optimal values for the QC acceptance criteria were assessed using ROC curves and calculating the Youden index for both PTV and OAR measurement points (see also section 2.9).

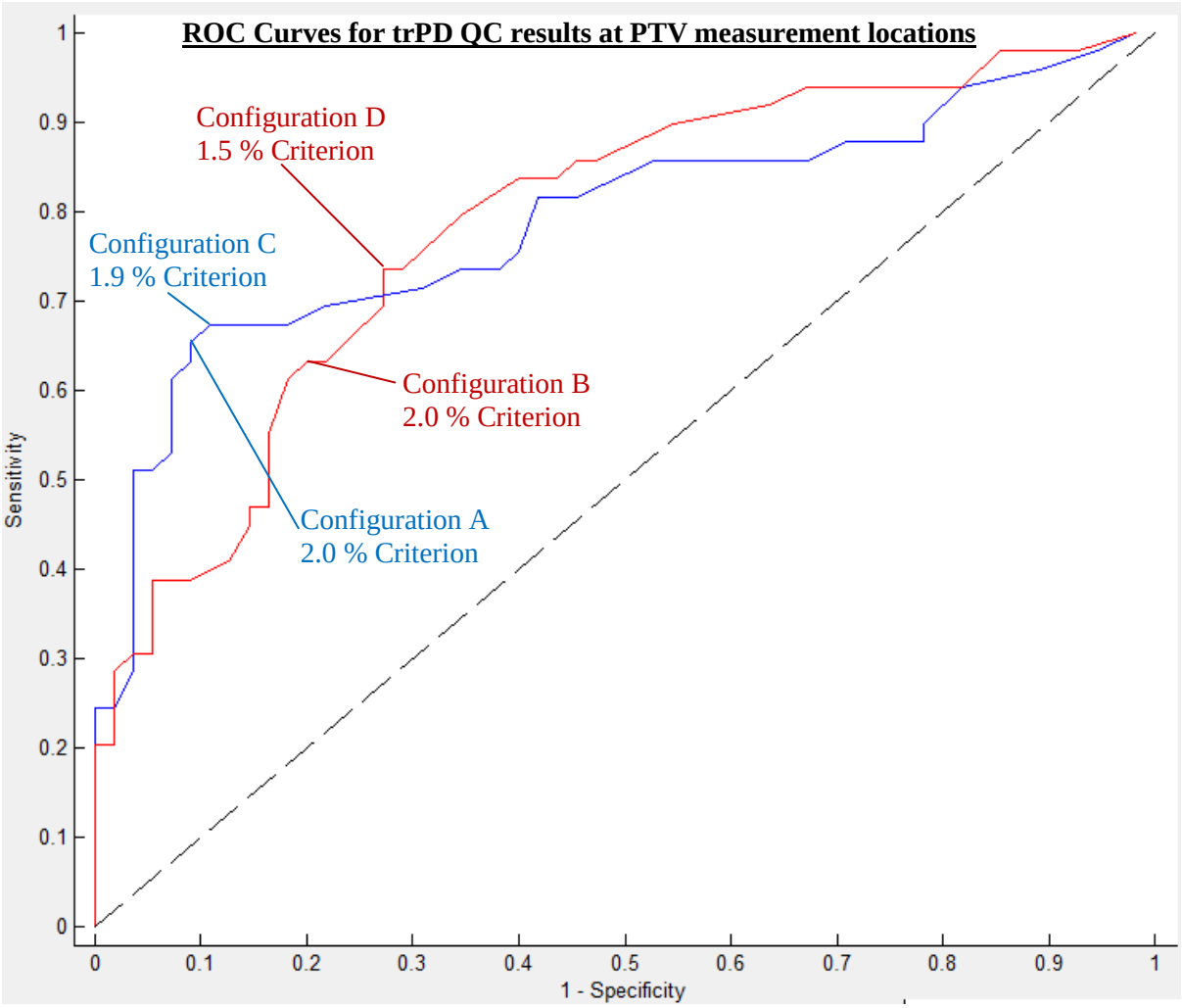

Figure 3.2: *ROC curves for trPD verification measurements at the PTV points of interest for both the clinical beam model (blue line) and the adjusted beam model (red line). The black dashed line represents the 0.5 area under the curve (AUC) value. The positions on the curve which correspond to configurations **A** and **B** (2.0% passing criterion) as well as the optimal acceptance criterion for each curve (configurations **C** and **D**, see **Table 3.7**) are also indicated.*

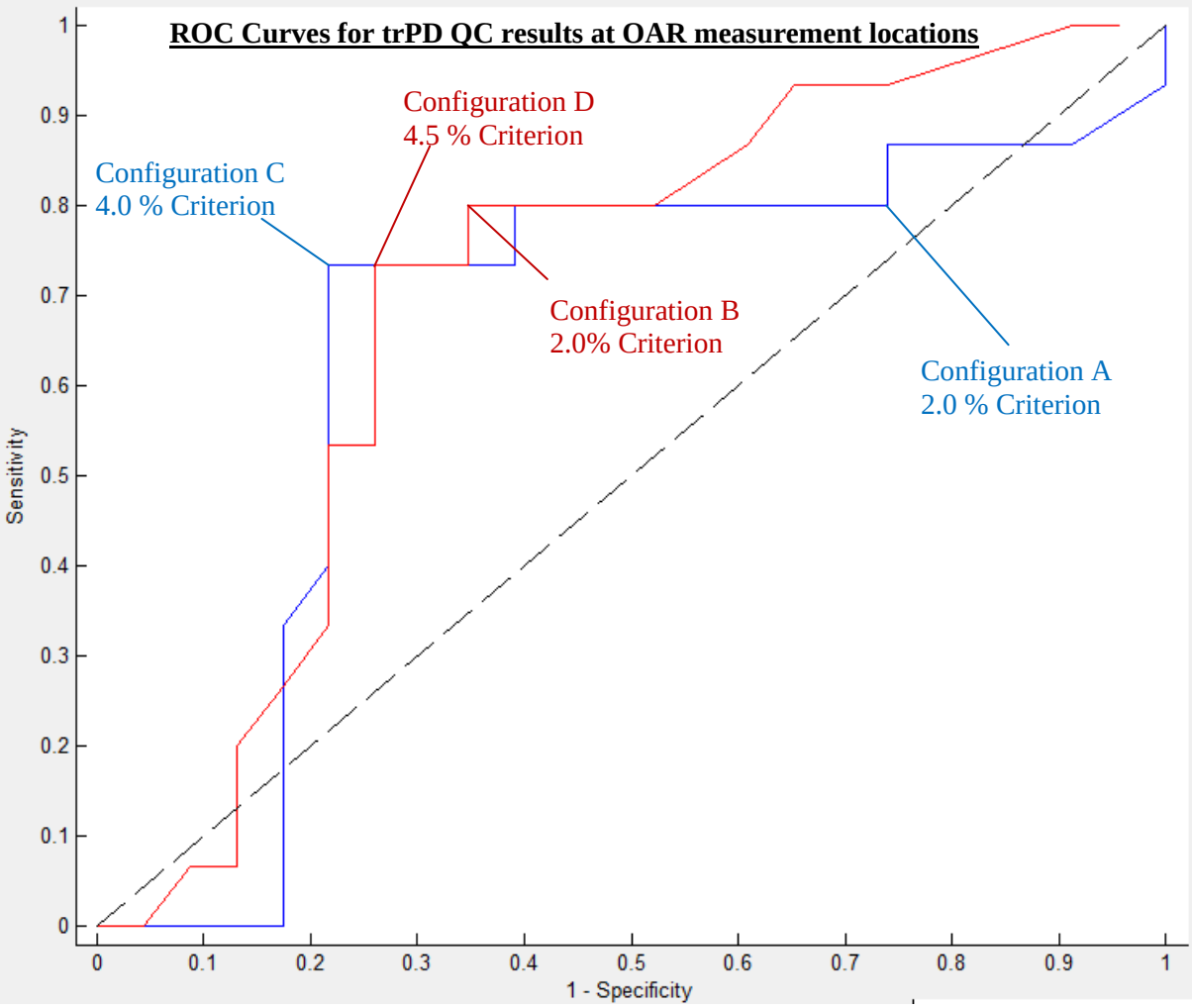

Figure 3.3: *ROC curves for trPD verification measurements at the OAR points of interest for both the clinical beam model (blue line) and the adjusted beam model (red line).*

ROC curves are shown above for the trPD QC results for both the clinical beam model (blue) and adjusted beam model (red) for measurements made at both the PTV (**Figure 3.2**) and OAR (**Figure 3.3**) locations. The optimal values for S_1 and Sp_1 as indicated by the Youden index are included in **Table 3.7**, along with the AUC and optimal QC threshold for the integral point dose QC method.

Table 3.7: *Metrics characterising the efficiency of the integral point dose measurements for both the PTV and OAR measurement locations. S_1 and Sp_1 values are those determined using configuration **C** and **D**. The values in brackets represent the change in a given result from configurations **A** and **B** respectively.*

PTV	Configuration C (change from A)	Configuration D (change from B)
AUC	0.79	0.79
S_1	67.3% (+2.0%)	73.5% (+10.2%)
Sp_1	89.1% (-1.8%)	72.7% (-7.3%)
Optimal QC acceptance criterion	±1.9% (-0.1%)	±1.5% (-0.5%)

OAR	Configuration C (change from A)	Configuration D (change from B)
AUC	0.65	0.70
S_1	73.3% (-6.7%)	73.3% (-6.7%)
Sp_1	78.3% (+52.2%)	73.9% (+8.7%)
Optimal QC acceptance criterion	±4.0% (+2.0%)	±4.5% (+2.5%)

Optimisation of the QC acceptance criterion only resulted in a small change in S_1 and Sp_1 values for the PTV locations. A large increase in $Sp_1$was observed for the OAR measurement locations using the clinical beam model.

Inclusion of OAR specific MLC intentional error plans into the set of PTV point verification measurements could potentially be another cause for the unexpected low sensitivity values that were obtained. It is not unreasonable to expect that a point dose measurement with the detector at a location that is reasonably far away from the location where the intentional error will impact, will not be detected (for example, mean distance between PTV point and SC point was 6.6 cm, while mean PTV to chiasm distance was 5.2 cm). Therefore, the ROC analysis was repeated while excluding all OAR specific intentional error plan measurements.

ROC Curves for trPD QC results at PTV measurement locations excluding OAR specific MLC errors

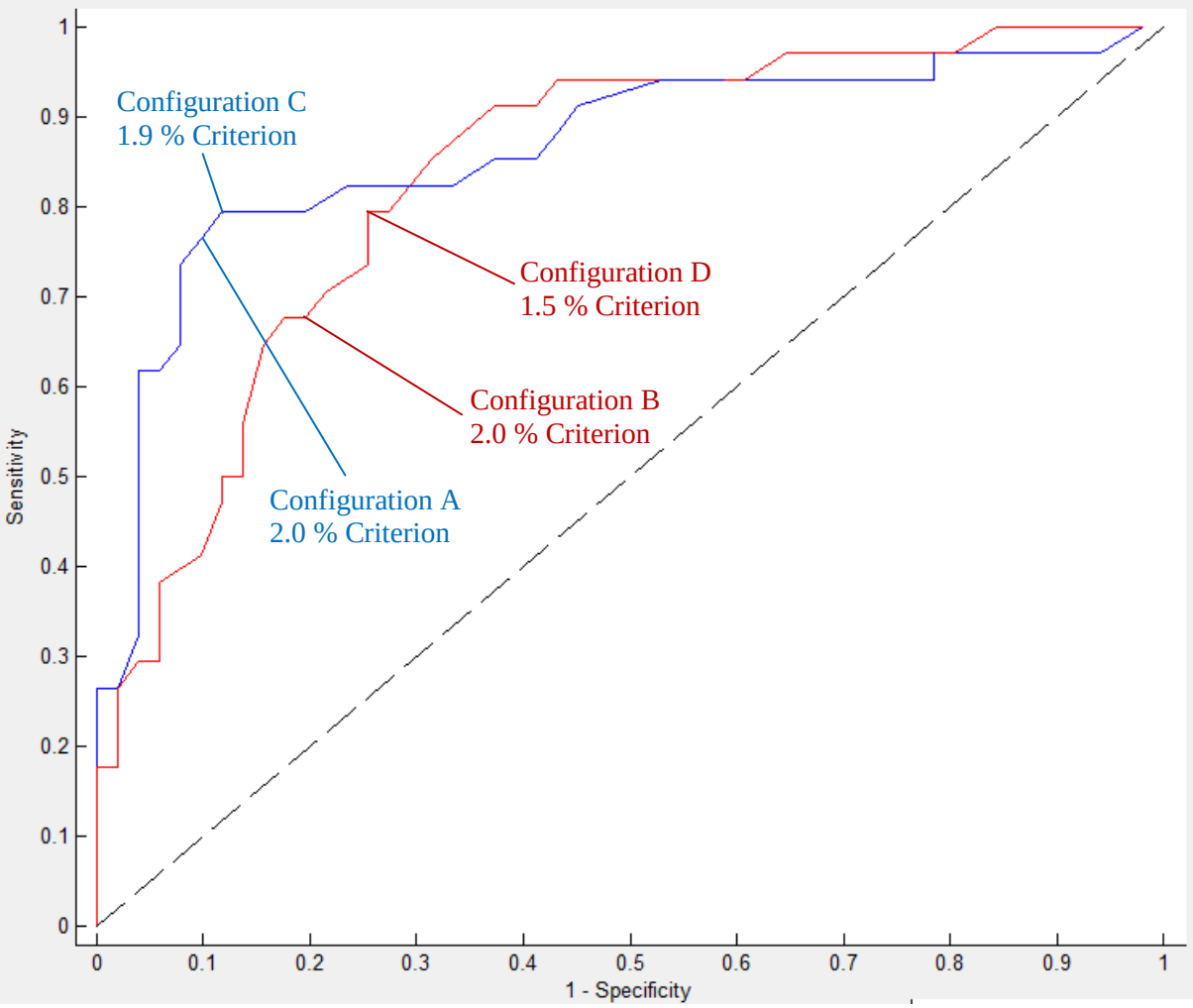

Figure 3.4: *ROC curves for trPD results excluding MLC shift errors near specific OARs for both the clinical beam model (blue line) and the adjusted beam model (red line).*

When these four plans are excluded, the sensitivity increases noticeably with only a small reduction in specificity (**Table 3.8**).

Table 3.8: *AUC, S_1, Sp_1 (as well as change in S_1 and Sp_1), and the optimal QC threshold for the point dose method based on the ROC curves in* ***Figure 3.4.*** *S_1 and Sp_1 values are those determined using for configurations* ***C*** *and* ***D****.*

PTV	Configuration C (change from A)	Configuration D (change from B)
AUC	0.86	0.83
S_1	79.4% (+14.1%)	79.4% (+16.1%)
Sp_1	88.2% (-2.7%)	74.5% (-5.5%)
Optimal QC acceptance criterion	±1.9% (-0.1%)	±1.5% (-0.5%)

This post-hoc analysis highlights the fundamental limitation of point dose measurements; they only measure dose at a single point. This limitation is further discussed in section 4.5.1.

3.2.2. S_2

S_2 was calculated as per Equation 2.9 for plans containing introduced delivery errors to determine the sensitivity of trPD measurements as the ratio of the change in output over the change in input. The median S_2 value was approximately 100% for the majority of error modes (**Table 3.9** and **Table 3.10**). However some S_2 values were noticeably different for 6 error modes for which part of the results were outside the 95th percentile of all data. For these error modes either the change in input was very small (1 mm MLC translation error), the detector was at a location well away from the point where the error impacted the 3D dose distribution (all MLC shift errors near a specific OAR), or two different errors cancelled each other out (plan containing both MU 3% high and MLC 1 mm closed shift errors). Specifically for OAR specific MLC errors it was found that the values for S_2 at OAR locations were closer to 100% compared to the measurements at the PTV location.

Table 3.9: *Median and range of S_2 values over five patients for each plan containing introduced errors for measurements made at the PTV location. Values in the table for which S_2 exceeded ±36% (95th percentile of all data) of 100% (median value) are shaded orange.*

Plan error(s)	Median S_2	S_2 Range
MU 3% low	102.8%	96.6% - 103.8%
MU 1.5% low	107.4%	93.1% - 119.6%
MU 1.5% high	93.4%	92.9% - 107.6%
MU 3% high	101.3%	93.7% - 103.7%
Output w gantry angle 8%	102.8%	97.9% - 115.9%
Output w gantry angle 4%	106.3%	89.7% - 117.8%
MLC 1 mm closed	99.0%	94.7% - 111.9%
MLC 0.5 mm closed	102.0%	93.1% - 105.5%
MLC 0.5 mm open	96.3%	92.1% - 106.6%
MLC 1 mm open	102.9%	94.0% - 111.0%
MLC 1 mm translation	119.6%	75.6% - 416.2%
MLC SC 1 mm	73.3%	46.8% - 111.0%
MLC SC 2 mm	109.6%	92.2% - 139.2%
MLC BS 2 mm	105.0%	97.3% - 146.2%
MLC Chiasm 2 mm	21.7%	0.0% - 121.7%
MU 3% high, MLC 1 mm closed	83.3%	-6.4% - 123.8%

Table 3.10: *Median S_2 and range of S_2 for plans containing the introduced MLC shifts near a given OAR when conducting trPD measurements at the specified OAR location.*

Plan error	Median S_2	S_2 Range
MLC SC 1 mm	99.6%	97.2% - 100.7%
MLC SC 2 mm	99.8%	93.1% - 107.8%
MLC BS 2 mm	94.7%	87.1% - 106.2%
MLC Chiasm 2 mm	80.6%	77.1% - 82.3%

3.2.3. S_3

S_3 was calculated as per Equation 2.11 for plans containing introduced errors to determine the sensitivity of trPD measurements as the ratio of the change in QC result over the change in error magnitude. For systematic error modes (MU errors, output variation with gantry angle errors and MLC open/closed shift errors) the median values for S_3 are very similar regardless of error magnitude (**Table 3.11** and **Table 3.12**). There is larger variability in OAR specific MLC shift errors, again due to the PTV measurement location being out of the volume affected by these error modes. S_3 values for OAR specific MLC errors measured at the PTV are much lower compared to the corresponding measurements at the OAR location.

Table 3.11: *Median and range of S_3 values for each individual error type measured at the PTV location using the trPD method for plans calculated using the clinical beam model.*

Plan error	Unit	Median S_3	S_3 Range
MU 3% low	$\%.\%^{-1}$	1.0	1.0 - 1.1
MU 1.5% low	$\%.\%^{-1}$	1.0	0.9 - 1.1
MU 1.5% high	$\%.\%^{-1}$	1.0	0.7 - 1.1
MU 3% high	$\%.\%^{-1}$	1.0	1.0 - 1.0
Output w gantry angle 8%	$\%.\%^{-1}$	0.3	0.3 - 0.3
Output w gantry angle 4%	$\%.\%^{-1}$	0.3	0.3 - 0.4
MLC 1 mm closed	$\%.mm^{-1}$	3.5	1.7 - 4.5
MLC 0.5 mm closed	$\%.mm^{-1}$	3.6	2.4 - 5.2
MLC 0.5 mm open	$\%.mm^{-1}$	3.8	2.6 - 4.6
MLC 1 mm open	$\%.mm^{-1}$	3.8	2.6 - 4.7
MLC 1 mm translation	$\%.mm^{-1}$	-0.6	-0.9 - 0.6
MLC SC 1 mm	$\%.mm^{-1}$	0.2	-0.3 - 0.7
MLC SC 2 mm	$\%.mm^{-1}$	0.2	0.2 - 0.8
MLC BS 2 mm	$\%.mm^{-1}$	2.4	1.5 - 3.0
MLC Chiasm 2 mm	$\%.mm^{-1}$	0.0	0.0 - 0.1
DLG 1.2 mm	$\%.mm^{-1}$	2.0	-0.8 - 7.0
ETSS X 1.5 mm	$\%.mm^{-1}$	-0.4	-1.3 - 0.7

Table 3.12: *Median and range of S_3 values for each individual error type measured at the OAR location using the trPD method for plans calculated using the clinical beam model.*

Plan error	Unit	Median S_3	S_3 Range		
MLC SC 1 mm	%.mm^{-1}	3.6	2.9	-	4.0
MLC SC 2 mm	%.mm^{-1}	3.8	3.2	-	4.3
MLC BS 2 mm	%.mm^{-1}	5.3	2.4	-	8.3
MLC Chiasm 2 mm	%.mm^{-1}	6.4	3.9	-	11.6

The trPD QC results for the MU errors and systematic MLC shift errors were plotted as a function of the error magnitude in **Figure 3.5** and **Figure 3.6** respectively to further illustrate the results obtained for S_3.

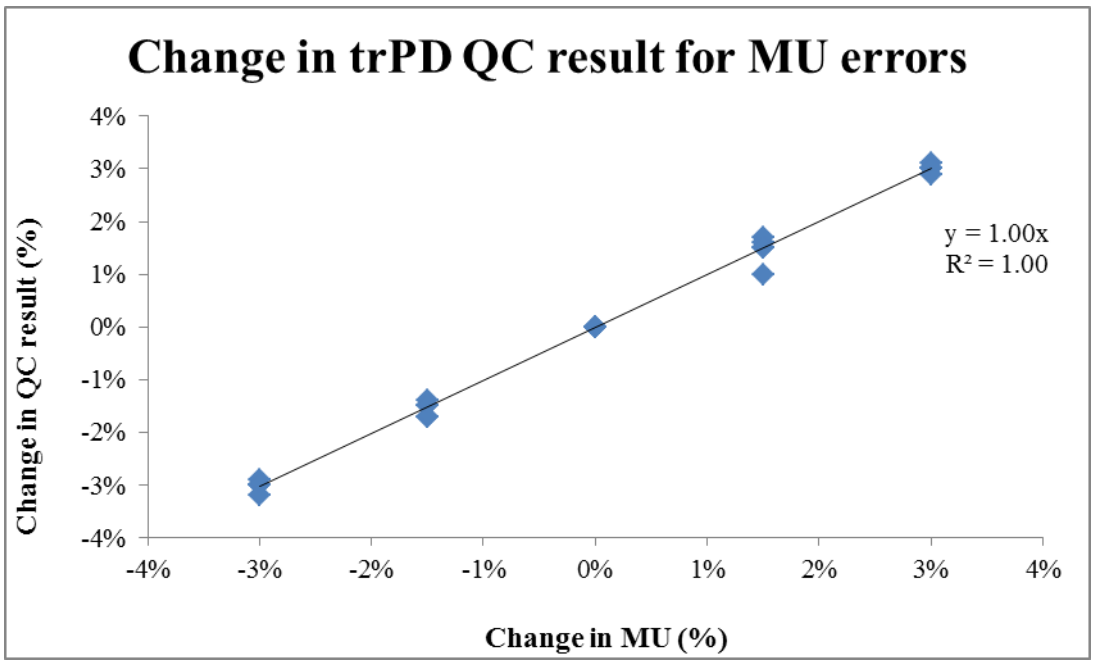

Figure 3.5: *Change in trPD QC results for systematic MU errors for PTV measurement points. Markers represent each individual QC measurement and the line represents the linear regression trend line over all data.*

The linear regression of QC result against change in MU resulted in a strong correlation (R^2 = 1.00) and trend line with a slope of 1.00. Thus, any percentage change in MU will nearly always result in the same percentage change in trPD measured dose. Furthermore, if the current 2.0% QC acceptance criterion was applied, only MU errors larger than ±2.0% will be detected by the trPD technique.

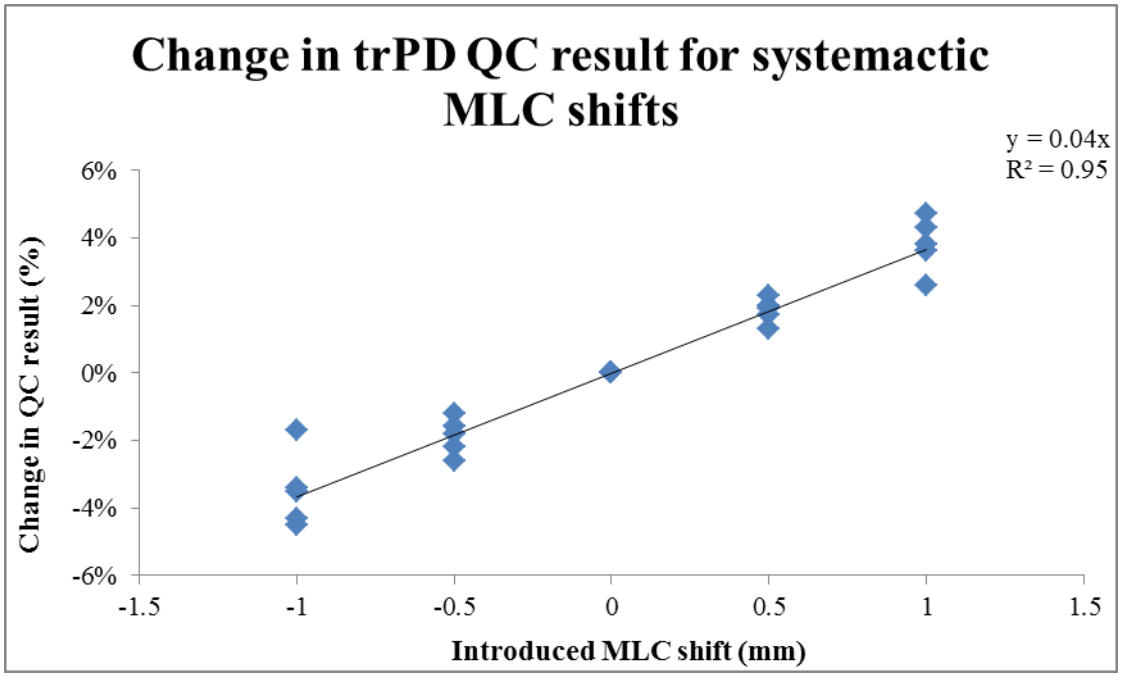

Figure 3.6: *Change in trPD QC results for systematic MLC shift errors for PTV measurement points. Markers represent each individual QC measurement and the line represents the linear regression trend line over all data.*

The linear regression of trPD QC results against change in magnitude of systematic MLC shift also showed a strong correlation (R^2 = 0.95). However there was more inter-patient variation compared to the plot for MU errors. The relationship between trPD result and change in systematic MLC shift was linear, with a slope of 4 %.mm^{-1}; indicating that a systematic opening or closing of the MLC bank will result in a change in dose measured using the trPD method by 4 % per mm of the MLC bank shift. Thus, with the current 2.0% QC passing criterion, any systematic MLC open or closed shift less than 0.5 mm is not likely be detected, and any shift greater than 0.5 mm is likely to be detected.

3.2.4. Sp_2

Specificity representing the ability of the trPD technique to resolve and identify different error modes was assessed using the methodology outlined in section 2.8.2. Due to time constraints, this initial feasibility study was limited to PTV measurement points and the inclusion of either an MLC or an MU error. Furthermore, this analysis was conducted using the clinical beam model, because the highest S_1 and Sp_1 values were obtained for this model (see section 3.2.1).

Analysis of Results Averaged over Multiple Patients

As a first step, an inventory was made of the impact of specific error types on the observed deviations averaged over all patients. Previous experience with trPD measurements had shown that the results averaged over a large number of measurements enabled identification of the cause of systematic deviations, while individual measurements displayed a considerably smaller signal to noise ratio (SNR). Results were plotted for the 5 error-free plans (**Figure 3.7**). This plot confirms the earlier observation that the SNR is considerably lower for individual patients compared to the average over 5 patients, particularly in region II. The results for the error-free plans compared to plans including an intentional 3% MU increase or decrease were plotted (**Figure 3.8**). In this figure the three plots diverge for CPs where the detector is not behind the MLCs (DTFE > 0 cm). The results for the error-free plans compared to plans including an intentional 1 mm MLC open or closed shifts were also plotted (**Figure 3.9**). This figure shows that the MLC open shift shows a positive deviation and the MLC closed shift error a negative deviation in region II compared to the error-free plans, while in region III all three lines converge to zero again.

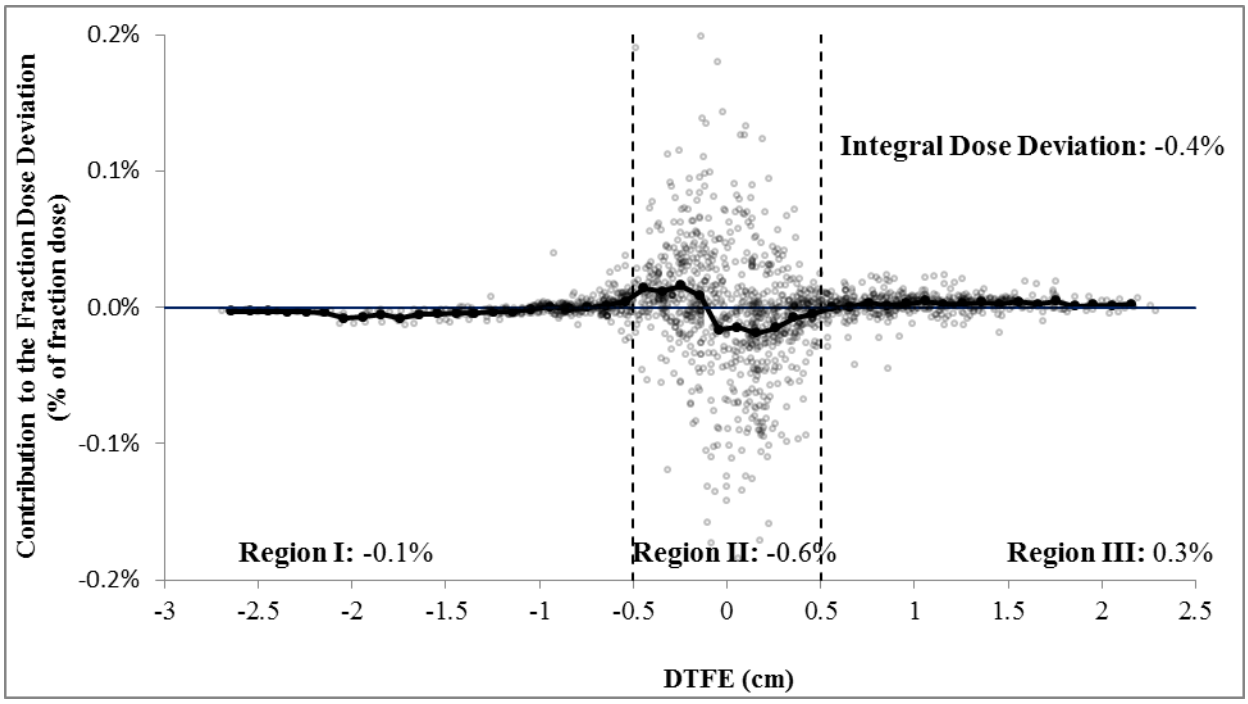

Figure 3.7: *Dose deviation per CP against DTFE for the five* error-free *treatment plans, indicating the integrated deviation per region averaged over all patients. The markers represent the individual deviation per CP for each individual patient and the black line represents the average contribution to the fraction dose deviation over 5 plans for a given DTFE value. The vertical dashed lines represent a DTFE value of ±0.5 cm which was used to define the three regions.*

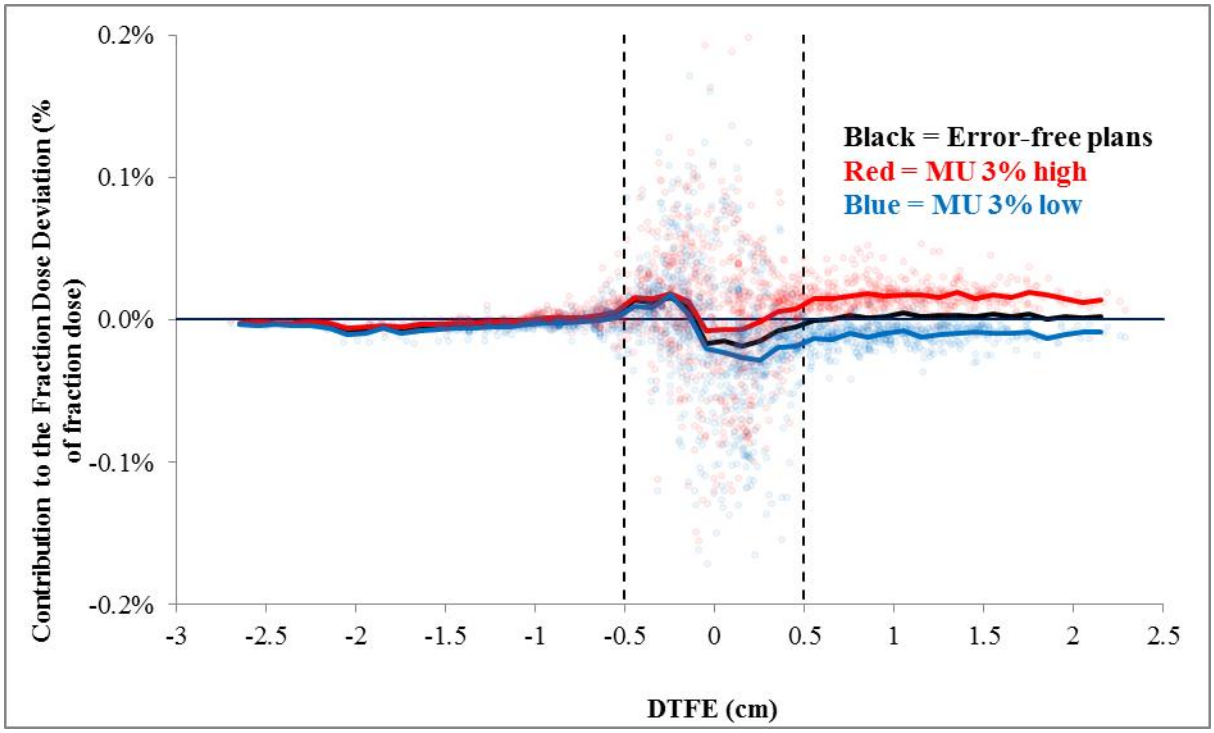

Figure 3.8: *Dose deviation per CP against DTFE for the error-free treatment plans (black), 3% MU increase (red), and the 3% MU decrease (blue). Solid lines represent the average contribution to fraction dose deviation for a given DTFE value while markers represent the individual deviation per CP for each individual patient.*

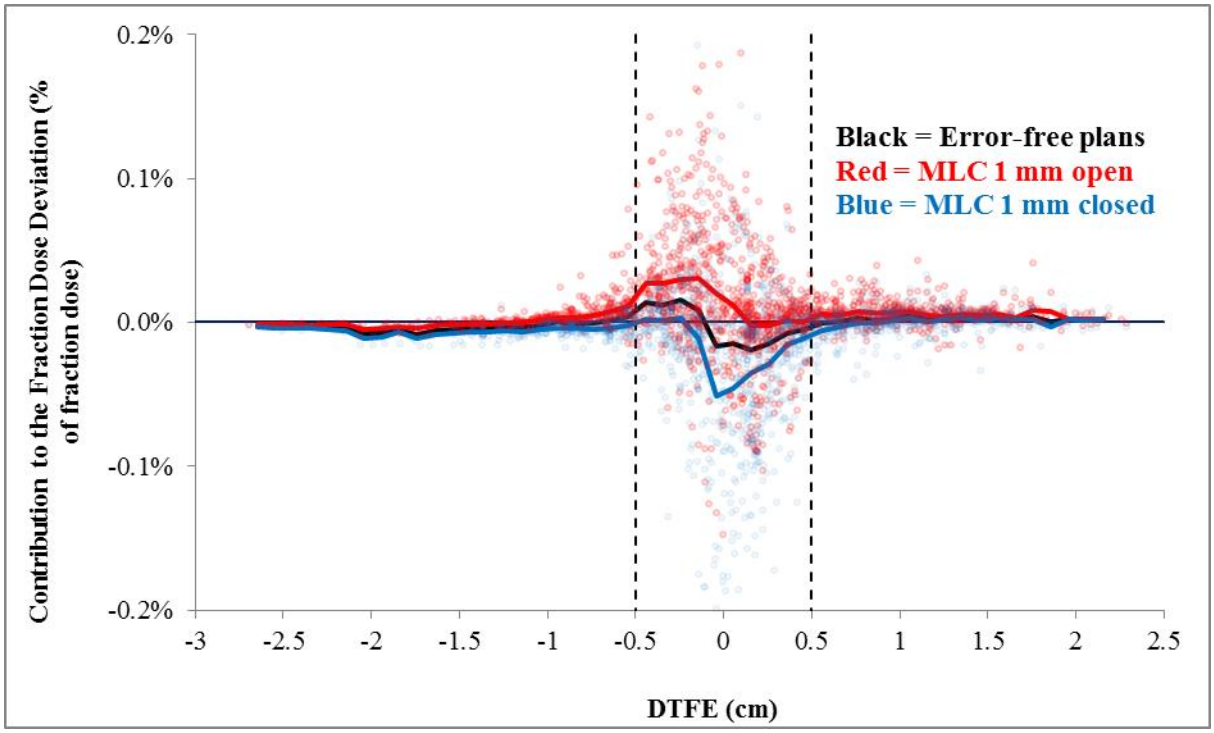

Figure 3.9: *Dose deviation per CP against DTFE for the error-free treatment plans (black), 1mm MLC open shift (red), and the 1mm MLC closed shift (blue). Solid lines represent the average contribution to fraction dose deviation for a given DTFE value.*

Figure 3.7 - 3.9 show that MLC shift errors and MU errors result in distinct deviations for regions II and III which can potentially be used to resolve different error types.

Based on the observed change in deviations and the corresponding variance for systematic MLC errors and MU errors, detection criteria for these error types were defined as follows:

a) MU errors:

$$|\Delta_{\text{mean}}^{\text{III}}| > TH \qquad \text{Equation 3.1a}$$

$$\left|\Delta_{\text{mean}}^{\text{II, corr}}\right| < TH \qquad \text{Equation 3.1b}$$

$$CI_{95\%}^{mean,\ \text{III}} \ni 0 \qquad \text{Equation 3.1c}$$

b) MLC errors:

$$\left|\Delta_{\text{mean}}^{\text{II, corr}}\right| > TH \qquad \text{Equation 3.2a}$$

$$|\Delta_{\text{mean}}^{\text{III}}| < TH \qquad \text{Equation 3.2b}$$

Where:

- $\Delta_{\text{mean}}^{\text{X}}$ represents the average contribution to the overall fraction deviation in DTFE region X relative to the baseline deviation
- CI is the corresponding 95% confidence interval for region III. This baseline deviation was determined by calculating the average contribution to the fraction deviation as a function of the DTFE for all error-free verification plans
- *TH* is the threshold that needs to be determined for an optimal specificity

For the detection of MLC errors using region II, a correction is applied to the mean deviation in region II (resulting in $\Delta_{\text{mean}}^{\text{II, corr}}$). This takes into account the impact of potential MU changes to the average deviation in region II[b], which is defined as the part of region II where DTFE > 0.

$$\left|\Delta_{\text{mean}}^{\text{II, corr}}\right| = \frac{N^{\text{IIa}}.|\Delta_{\text{mean}}^{\text{IIa}}|+N^{\text{IIb}}.(|\Delta_{\text{mean}}^{\text{IIb}}|-|\Delta_{\text{mean}}^{\text{III}}|)}{N^{\text{IIa}}+N^{\text{IIb}}} \qquad \text{Equation 3.3}$$

Where:

- N^{X} is the number of data points in region X.

No detection criterion including a 95% CI for region II was defined for MLC errors because it was anticipated this would not be an effective addition to the threshold criterion due to the low SNR in region II (see **Figure 3.7 - 3.9**).

Sp_2 for Individual Patient Results

After application of Equations 3.1 – 3.3 to the results of the individual verification measurements, it was found that a maximum specificity was obtained using *TH* = 0.0075. With this threshold level, all MU errors of ± 3% and MLC errors of ± 1 mm were correctly identified, and no errors were incorrectly identified (**Table 3.13**). In addition, 50% of the smaller MU errors and 60% of the smaller MLC errors were correctly identified.

Table 3.13 *Detected error types for the individual verification plans using a threshold TH = 0.0075.*

	Patient 1	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**	**Patient 1**	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**
Plan error(s)	*Equations 3.1a-c*					*Equations 3.2a-b*				
No error	-	-	-	-	-	-	-	-	-	-
MU 3 % low	MU	MU	MU	MU	MU	-	-	-	-	-
MU 1.5% low	-	-	-	MU	-	-	-	-	-	-
MU 1.5% high	MU	-	MU	MU	MU	-	-	-	-	-
MU 3% high	MU	MU	MU	MU	MU	-	-	-	-	-
MLC closed 1 mm	-	-	-	-	-	MLC	MLC	MLC	MLC	MLC
MLC closed 0.5 mm	-	-	-	-	-	MLC	MLC	-	MLC	MLC
MLC open 0.5 mm	-	-	-	-	-	-	MLC	MLC	-	-
MLC open 1 mm	-	-	-	-	-	MLC	MLC	MLC	MLC	MLC

3.3. Film Results

Film dosimetry for all error-free plans was completed during 9 separate measurement sessions. For each patient the PTV plane was measured at least twice, while the OAR planes were measured once or twice depending on which plans containing introduced errors were measured during that session (for instance, a measurement of the error-free plan SC plane was measured if plans containing introduced errors for the SC were to be measured in that session). These results are included in **Table 3.14**.

Table 3.14: *Film results for measurement of the error-free plans. All results are based on configuration **A**. $P_\gamma^{\{2\%,2mm\}}$ represents the observed pass rate using a {2%; 2mm} γ-criterion. False Positive results are bolded and highlighted in orange; all other results are TNs.*

Patient	Plane of Interest	Measurement 1 $P_\gamma^{\{2\%,2mm\}}$ (%)	Measurement 2 $P_\gamma^{\{2\%,2mm\}}$ (%)	Measurement 3 $P_\gamma^{\{2\%,2mm\}}$ (%)
Patient 1	PTV	96.9%	97.3%	**68.9%**
	SC PRV	98.0%	98.7%	-
	BS PRV	94.6%	-	-
	Chiasm	-	-	-
Patient 2	PTV	96.4%	93.5%	95.5%
	SC PRV/ BS PRV	**78.6%**	92.8%	-
	Chiasm	86.1%	-	-
Patient 3	PTV	90.1%	96.7%	-
	SC PRV/ BS PRV	**65.1%**	**83.6%**	-
	Chiasm	88.3%	-	-
Patient 4	PTV	**80.1%**	95.8%	-
	SC PRV/ BS PRV	**76.7%**	92.9%	-
	Chiasm	**82.6%**	-	-
Patient 5	PTV	85.5%	87.3%	-
	SC PRV/ BS PRV	**79.6%**	**68.7%**	-
	Chiasm	85.9%	-	-

All films located at the PTV plane resulted in true negatives except for two cases (patient 1 third measurement and patient 4 first measurement), indicating that the specificity for PTV measurements seems to be high. In contrast, 8 TNs and 7 FPs in total were observed for film measurements at the OAR locations. Concurrent to this study, a number of routine patient-specific QC measurements using film were resulting in low γ pass rates. These occurrences showed a similar trend to the current study in that a reduced TNR for OAR planes and a few failing PTV planes were observed. A departmental review of the film dosimetry program at the WBCC showed that there was a large intra batch variation in film response for routine QC at WBCC (see section 3.3.2). This was also observed in the current study, as exemplified in **Figure 3.10**. These two films of the same verification plan

yielded a considerable difference in pass rate using a {2%; 2mm} γ-criterion: 80.1% and 95.8%, respectively.

This intra film batch variability could potentially account for a number of FP and FN results observed in this study and this effect is discussed further in section 4.5.2.

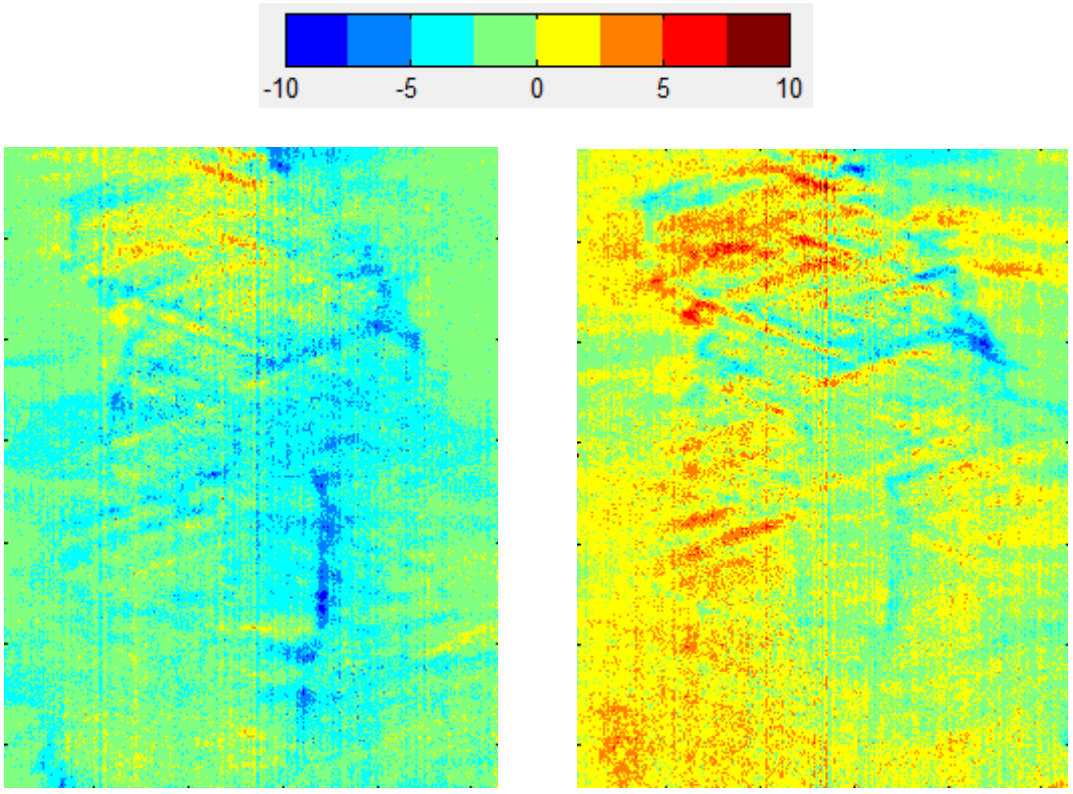

Figure 3.10: *Two PTV dose comparisons between measured film dose and TPS calculated dose for patient 4, where the films were measured on two separate occasions. The left hand image was the result from the first measurement session and the right hand image was the result from the second measurement session. Analysis was conducted using global dose differences (dose difference relative to the average PTV dose). The top colour bar defines the colour scale representing the percentage difference between measured and calculated dose at each pixel location.*

QC measurements using film dosimetry were carried out for all plans with intentional errors as described in section 2.6.3, and the results are summarised in **Table 3.15**.

Table 3.15: *QC results and indication of outcome for all plans containing introduced errors based on configuration **A**. Orange shading corresponds to false negatives and purple shading corresponds to false positives. All non-shaded results are either true negatives or true positives.*

Plan error(s)	Patient 1		Patient 2		Patient 3		Patient 4		Patient 5	
	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome
MU 3 % low	35.7%	TP	45.7%	TP	71.2%	TP	46.0%	TP	74.5%	TP
MU 1.5% low	**70.5%**	**FP**	**74.2%**	**FP**	94.4%	TN	**84.0%**	**FP**	96.1%	TN
MU 1.5% high	87.7%	TN	**80.9%**	**FP**	98.7%	TN	97.2%	TN	95.5%	TN
MU 3% high	58.2%	TP	77.2%	TP	**98.0%**	**FN**	**90.7%**	**FN**	72.5%	TP
Output w gantry angle 8%	50.2%	TP	68.5%	TP	**85.3%**	**FN**	70.2%	TP	83.5%	TP
Output w gantry angle 4%	97.6%	TN	97.6%	TN	87.6%	TN	**78.2%**	**FP**	88.0%	TN
MLC closed 1 mm	**86.0%**	**FN**	37.8%	TP	39.6%	TP	20.0%	TP	42.5%	TP
MLC closed 0.5 mm	95.1%	TN	90.7%	TN	83.5%	TP	67.6%	TP	**88.6%**	**FN**
MLC open 0.5 mm	87.8%	TN	73.1%	TP	**95.5%**	**FN**	**92.4%**	**FN**	93.3%	TN
MLC open 1 mm	47.1%	TP	78.8%	TP	84.7%	TP	**85.9%**	**FN**	75.8%	TP
MLC translation 1mm	97.4%	TN	95.0%	TN	91.8%	TN	92.6%	TN	97.3%	TN
MLC SC 1 mm - PTV	**93.0%**	**FN**	84.0%	TP	99.1%	TN	**98.6%**	**FN**	99.5%	TN
MLC SC 1 mm - SC	**98.5%**	**FN**	**90.5%**	**FN**	96.5%	TN	**85.6%**	**FN**	94.2%	TN
MLC SC 2 mm - PTV	83.7%	TP	78.6%	TP	**93.1%**	**FN**	74.7%	TP	**94.1%**	**FN**
MLC SC 2mm - SC	83.9%	TP	83.2%	TP	71.5%	TP	68.8%	TP	**89.9%**	**FN**
MLC BS 2 mm - PTV	**87.1%**	**FN**	**85.4%**	**FN**	**93.3%**	**FN**	83.1%	TP	**86.6%**	**FN**
MLC BS 2 mm - BS	**91.8%**	**FN**	**92.4%**	**FN**	**94.3%**	**FN**	83.2%	TP	**89.3%**	**FN**
MLC Chiasm 2 mm - PTV	-		**98.7%**	**FN**	99.5%	TN	**96.0%**	**FN**	94.1%	TN
MLC Chiasm 2 mm - Chiasm	-		**98.1%**	**FN**	99.1%	TN	**95.9%**	**FN**	93.1%	TN
MU 3% high, MLC 1mm closed	93.0%	TN	90.0%	TN	80.3%	TP	65.2%	TP	89.1%	TN
DLG 1.2 mm	92.2%	TN	99.0%	TN	**99.4%**	**FN**	95.6%	TN	95.9%	TN
ETSS X 1.5 mm	**71.5%**	**FP**	96.5%	TN	88.6%	TN	**80.5%**	**FP**	85.4%	TN

These results are summarised in truth tables comparing configurations **A** and **B** in **Table 3.16A - B** for PTV measurement planes and **Table 3.17A - B** for OAR measurement planes.

Table 3.16 A-B: *Truth tables based on the clinical relevance criteria as defined in section 2.2 for the film measurements at the PTV measurement planes.*

Configuration **A**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	30	19
Not clinically relevant	9	43

Configuration **B**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	32	17
Not clinically relevant	10	42

Table 3.17 A-B: *Truth tables based on the clinical relevance criteria as defined in section 2.2 for the film measurements at the OAR measurement planes.*

Configuration **A**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	5	10
Not clinically relevant	7	12

Configuration **B**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	9	6
Not clinically relevant	2	17

For PTV measurement planes, the truth tables for the film results are very similar to those obtained using the trPD method (see **Table 3.5A - B**) regardless of beam model used. Whereas for OAR film measurements, a higher number of both TNs and TPs were observed for the results for configuration **B** compared to configuration **A**.

3.3.1. S_1 and Sp_1

Plots of QC results (γ pass rate criterion only) for films located at the PTV plane against change in clinical relevance metrics were made to graphically illustrate the results for S_1 and Sp_1 (**Figure 3.11A - B**). Unlike the equivalent trPD plots (see **Figure 3.1A - B**), there was no linear correlation between the film γ pass rates and clinical relevance. When the change in both clinical relevance metrics is close to zero, most plans had a high γ pass rate. As the change in either clinical relevance metric increases, the γ pass rates begin to decrease albeit in a non-linear fashion. There were a number of points which did not follow this trend which resulted in FPs and FNs (total of 28 false results and 73 true results).

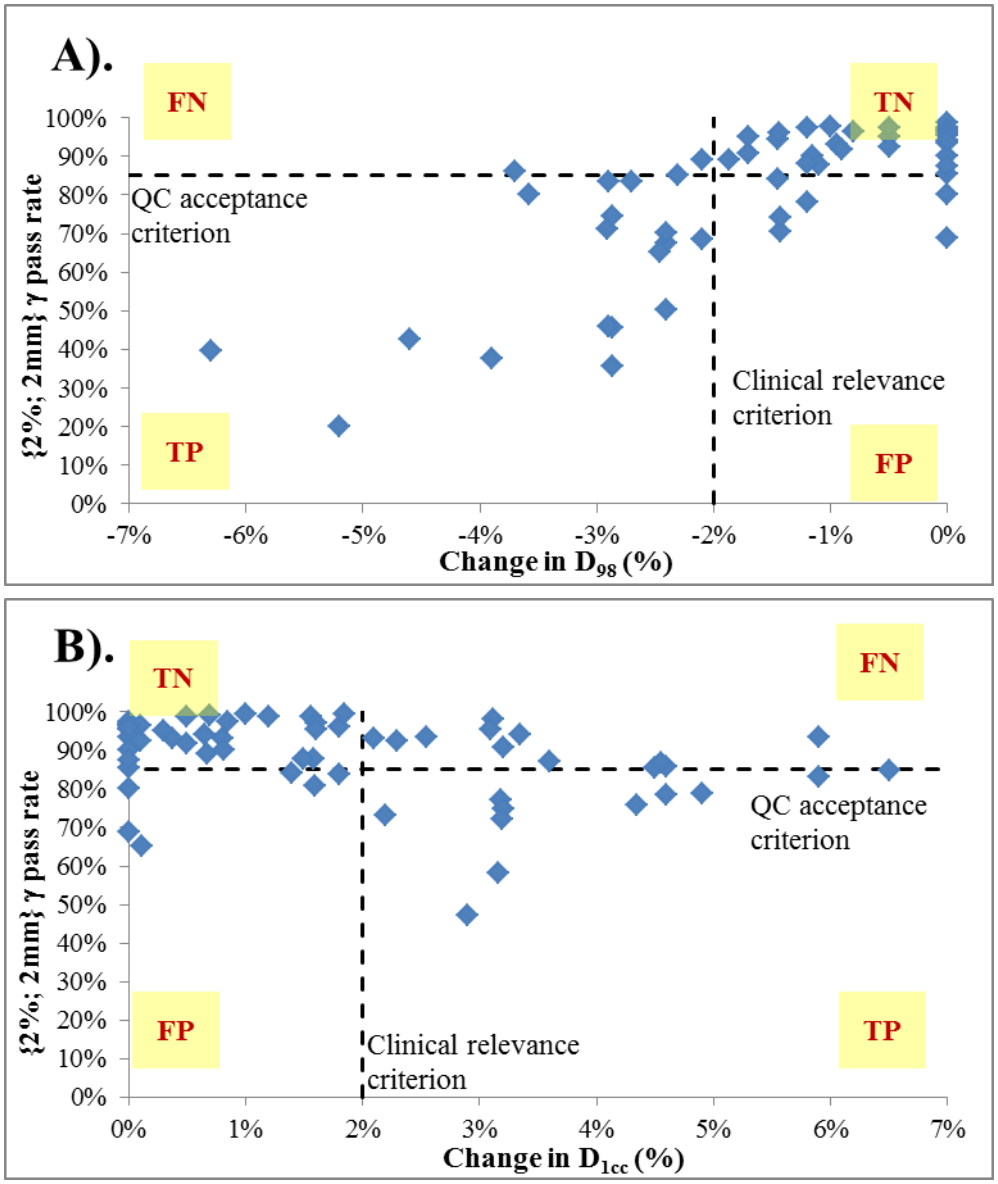

Figure 3.11 A (top) and B (bottom): *Plots of film results using configuration **A** against change in clinical relevance metrics with indications of QC passing criterion, clinical relevance criterion and regions corresponding to true and false positives and negatives.*

Subsequently, S_1 and Sp_1 metrics were calculated for PTV and OAR plane measurements after including all DVH metrics defining clinical relevance (**Table 3.1**) using Equation 2.8 and Equation 2.12 and these are given in **Table 3.18**.

Table 3.18: *S_1 and Sp_1 for the film method as determined using the methods outlined in sections 2.7.1 and 2.8.1 for both configuration **A** and **B** for both PTV measurement planes and OAR measurement planes.*

PTV	Configuration A	Configuration B
S_1	61.2%	65.3%
Sp_1	82.7%	80.8%

OAR	Configuration A	Configuration B
S_1	33.0%	60.0%
Sp_1	63.2%	89.5%

The sensitivity and specificity of the film measurements at PTV measurement planes was very similar to that obtained for the trPD analysis over both beam models (**Table 3.6**). Film measurements at OAR planes utilising configuration **B** resulted in similar S_1 and Sp_1 values to PTV measurements, while OAR measurements using configuration **A** had both a lower S_1 and Sp_1.

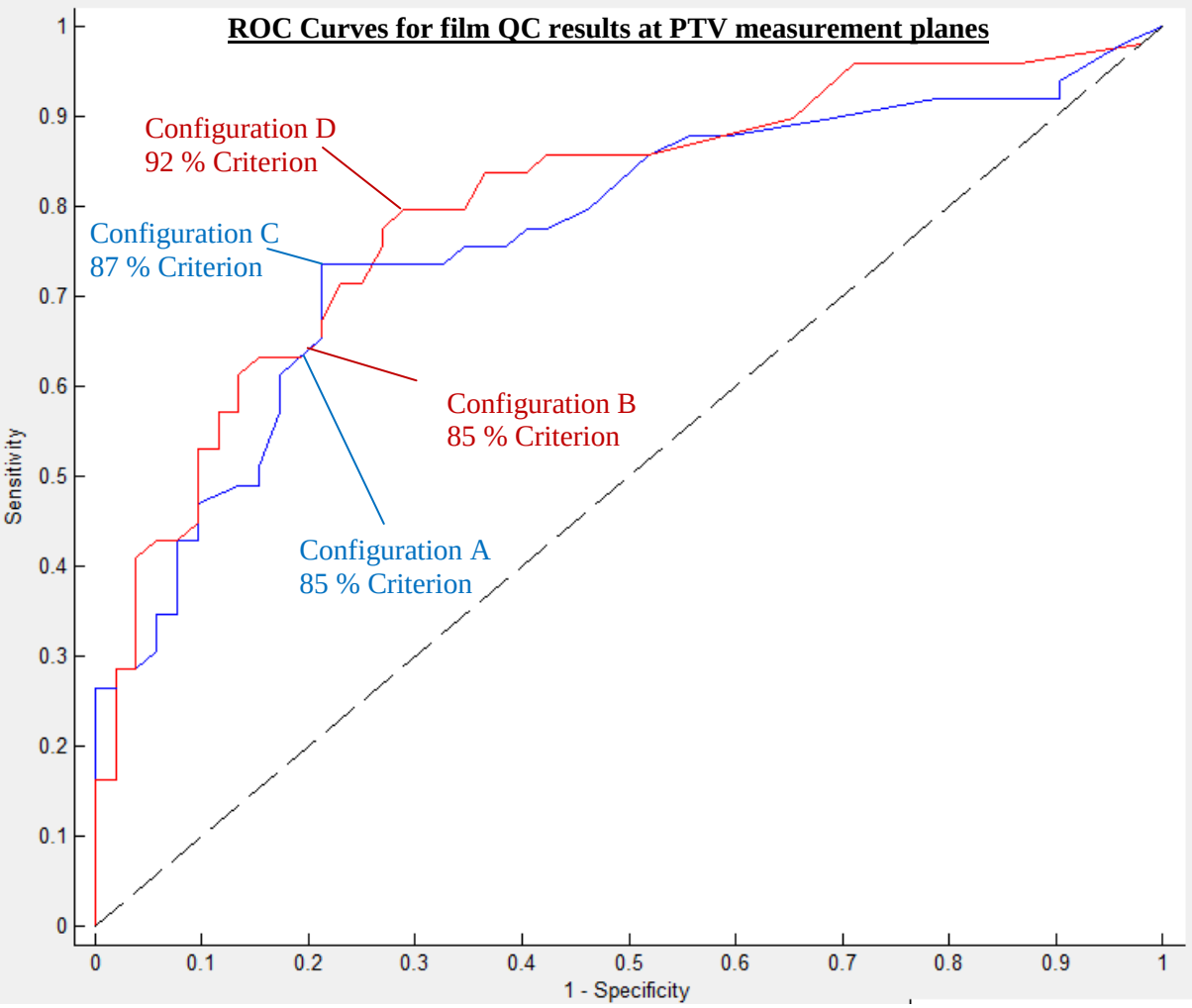

Figure 3.12: *ROC curves for film verification measurements at the PTV planes for both the clinical beam model (blue line) and the adjusted beam model (red line). The positions on the curves which correspond to configuration **A** and **B** (85% of points passing a {2%; 2mm} γ-criterion) as well as the optimal passing criterion for each curve (configuration **C** and **D**, see **Table 3.19**) are also indicated.*

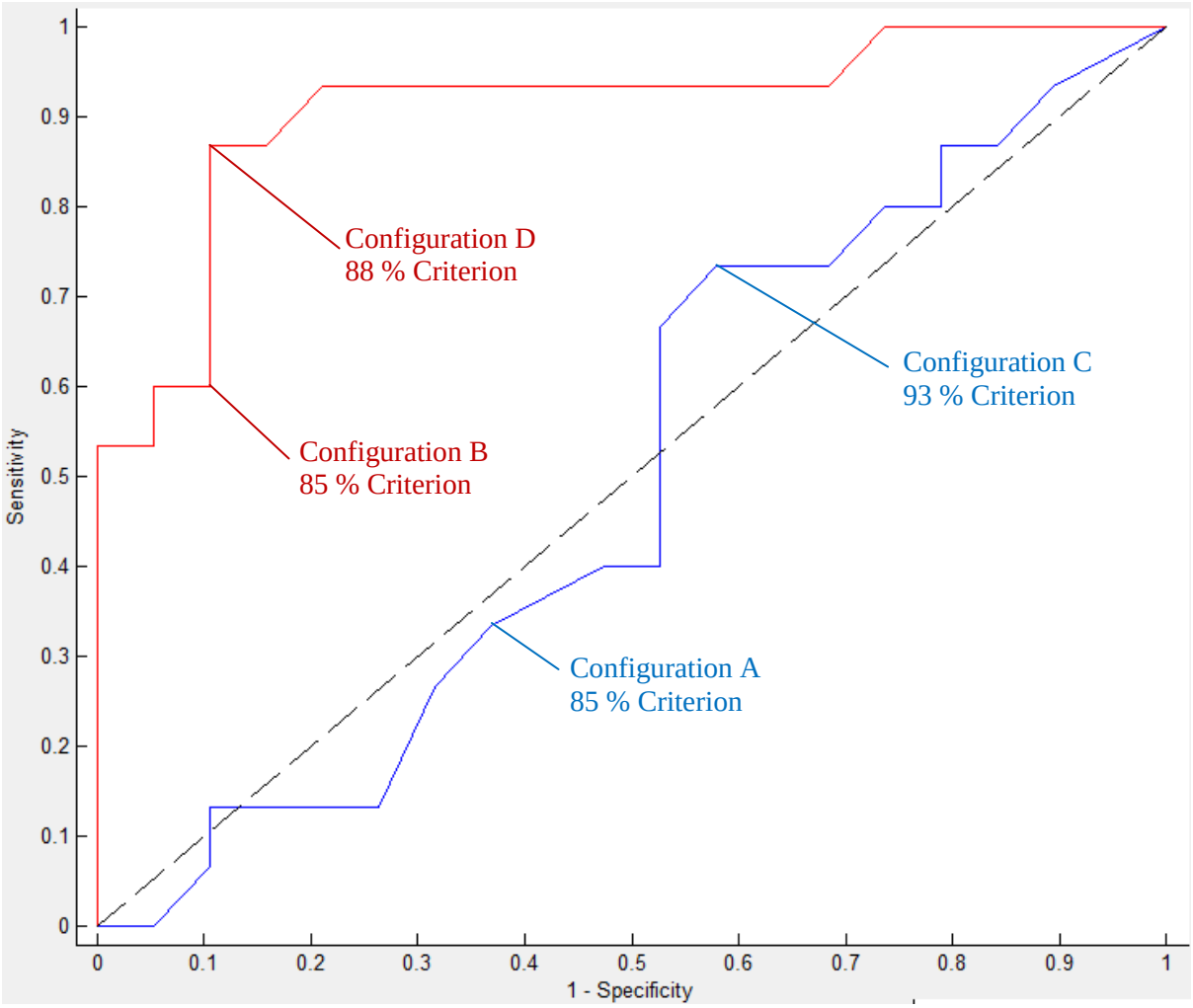

Figure 3.13: *ROC curves for film verification measurements at the OAR planes for both the clinical beam model (blue line) and the adjusted beam model (red line).*

Figure 3.12 and **Figure 3.13** show the ROC curves for the film QC results for both the clinical beam model (blue) and adjusted beam model (red) for the PTV and OAR measurement planes, respectively. The optimal values for S_1 and Sp_1 were determined using the Youden index. These values are included in **Table 3.19**, along with the AUC and optimal QC threshold for the film QC method.

Table 3.19: *Metrics characterising the efficiency of the film measurements for the PTV and OAR measurement planes. S_1 and Sp_1 values are those determined using configurations **C** and **D**. The values in brackets represent the change in a given result from configuration **A** and **B** respectively.*

PTV	Configuration C (change from A)	Configuration D (change from B)
AUC	0.77	0.80
S_1	73.5% (+11.4%)	79.6% (+14.3%)
Sp_1	78.8% (-3.9%)	71.2% (-9.6%)
Optimal QC acceptance criterion	87% (+2%)	92% (+7%)

OAR	Configuration C (change from A)	Configuration D (change from B)
AUC	0.50	0.91
S_1	73.3% (+40.0%)	86.7% (+26.7%)
Sp_1	42.1% (-21.1%)	89.5% (+0.0%)
Optimal QC acceptance criterion	93% (+8%)	88% (+3%)

Unlike the trPD analysis where S_1 increase for PTV but decreased for OAR locations, optimising the passing criterion led to overall improved sensitivity with only a small decrease in specificity. The exception to this trend was for the OAR measurement planes using the clinical beam model, where specificity was considerably reduced. It was not possible to obtain a criterion which had both S_1 and Sp_1 above 50%. This is reflected in the low AUC value of 0.50.

As for the trPD measurements, the OAR specific MLC errors may have led to a reduced number of TPs due to the negligible impact of these errors on the dose at the PTV measurement plane. To verify this, another set of ROC curves were generated excluding the OAR specific error plans (see **Figure 3.14**).

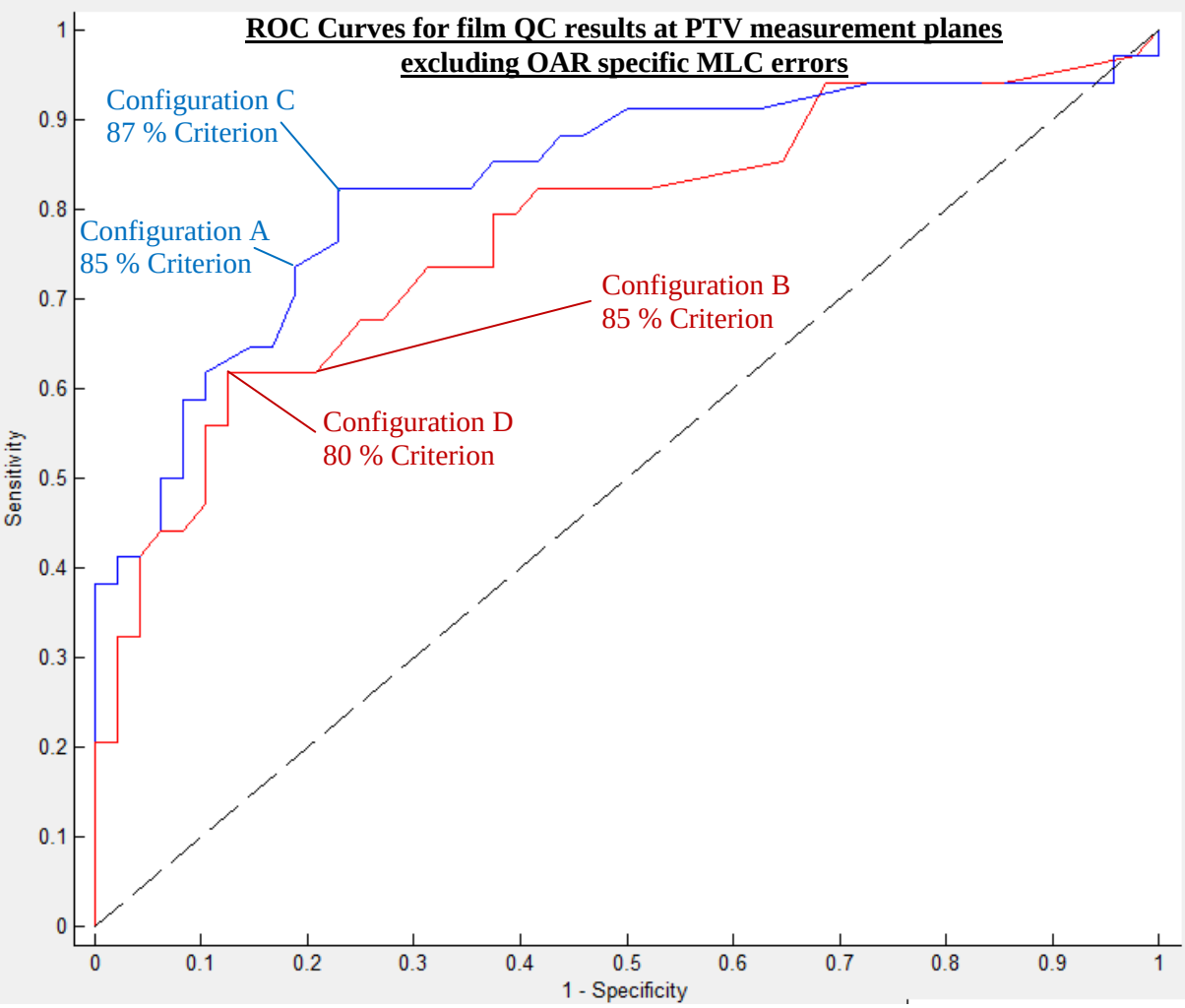

Figure 3.14: *ROC curves for film results excluding MLC shift errors near specific OARs for both the clinical beam model (blue line) and the adjusted beam model (red line) for PTV measurement planes only.*

With these four plans excluded, the AUC, S_1, Sp_1 and optimal cut off threshold were re determined and are given in **Table 3.20**.

Table 3.20: *AUC, S_1, Sp_1 (as well as change in S_1 and Sp_1), and the optimal QC threshold for the film method based on the ROC curves in **Figure 3.14** which excluded plans containing OAR specific errors. S_1 and Sp_1 values are those determined using configurations **C** and **D**.*

PTV	Configuration C (change from A)	Configuration D (change from B)
AUC	0.83	0.78
S_1	82.4% (+21.2%)	61.8% (-3.5%)
Sp_1	77.1% (-5.6%)	87.5% (+6.7%)
Optimal QC acceptance criterion	87% (+2%)	80% (-5%)

This analysis indicates that the film results at PTV measurement planes are slightly less efficient than the corresponding trPD measurements. This is reflected in the slight reduction of the film analysis AUC for both configuration **C** and **D** relative to the trPD results (0.03 for configuration **C**, 0.05 for configuration **D**, see **Table 3.8**).

3.3.2. S_2

In order to determine S_2 for the film dosimetry method, the current Matlab software would need to have been modified to allow the comparison of two film measurements. This would have required a reasonable amount of time to modify and test the software. Other data analysis was prioritised considering the time required to make this software change and therefore the S_2 film analysis could not be completed within the time constraints of this study.

3.3.3. S_3

S_3 was calculated as per Equation 2.10 for plans containing introduced errors to determine the sensitivity of film measurements as the ratio of the change in QC result over the change in error magnitude for plans calculated using the clinical beam model. Due to the non-linear behaviour of γ analysis, the values of S_3 for small error magnitudes were considerably smaller than for errors exceeding the {2%; 2mm} γ-criterion which functions as a threshold (see **Table 3.21** and **Table 3.22**). This is in contrast to the trPD results where S_3 displayed a linear behaviour.

For OAR measurement planes, the median S_3 value for all error modes is positive, indicating that γ pass rates increased for introduced error plans relative to the error-free plans.

Table 3.21: *Median S_3 and range of S_3 for each individual error type measured at the PTV plane using the film method for plans calculated using the clinical beam model.*

Plan error	Unit	Median S_3	S_3 Range
MU 3% low	$\%.\%^{-1}$	-11.0	-16.6 - -4.3
MU 1.5% low	$\%.\%^{-1}$	1.1	-14.2 - 5.9
MU 1.5% high	$\%.\%^{-1}$	1.3	-8.4 - 11.4
MU 3% high	$\%.\%^{-1}$	-4.4	-12.9 - 3.6
Output w gantry angle 8%	$\%.\%^{-1}$	-1.4	-3.4 - -0.5
Output w gantry angle 4%	$\%.\%^{-1}$	0.1	-2.3 - 1.0
MLC 1 mm closed	$\%.mm^{-1}$	-50.5	-60.1 - -10.8
MLC 0.5 mm closed	$\%.mm^{-1}$	-5.7	-26.3 - 3.3
MLC 0.5 mm open	$\%.mm^{-1}$	-2.5	-41.0 - 24.5
MLC 1 mm open	$\%.mm^{-1}$	-9.7	-49.8 - 5.8
MLC 1 mm translation	$\%.mm^{-1}$	1.5	-4.9 - 12.4
MLC SC 1 mm	$\%.mm^{-1}$	2.4	-9.5 - 18.5
MLC SC 2 mm	$\%.mm^{-1}$	-2.7	-8.9 - 4.3
MLC BS 2 mm	$\%.mm^{-1}$	0.6	-5.5 - 1.6
MLC Chiasm 2 mm	$\%.mm^{-1}$	4.5	1.1 - 8.0
DLG 1.2 mm	$\%.mm^{-1}$	-23.2	-38.6 - 11.7
ETSS X 1.5 mm	$\%.mm^{-1}$	-0.1	-1.8 - 1.0

Table 3.22: *Median and range of S_3 values for each individual error type measured at the OAR plane using the film method for plans calculated using the clinical beam model.*

Plan error	Unit	Median S_3	S_3 Range
MLC SC 1 mm	$\%.mm^{-1}$	8.9	-2.3 - 25.6
MLC SC 2 mm	$\%.mm^{-1}$	2.3	-7.0 - 5.2
MLC BS 2 mm	$\%.mm^{-1}$	3.2	-1.4 - 14.6
MLC Chiasm 2 mm	$\%.mm^{-1}$	5.7	3.6 - 6.7

Figure 3.15 and **Figure 3.16** illustrate the non-linear nature of S_3 for film dosimetry. By using a {2%; 2mm} γ-criterion that functions as a threshold, intentional errors with a smaller impact are not expected to result in lower γ pass rates. **Table 3.21**, **Figure 3.15** and **Figure 3.16** confirm that on average, this is the case. Note that the variation in the y-direction of both figures represents false positives/false negatives, which is noticeable as well.

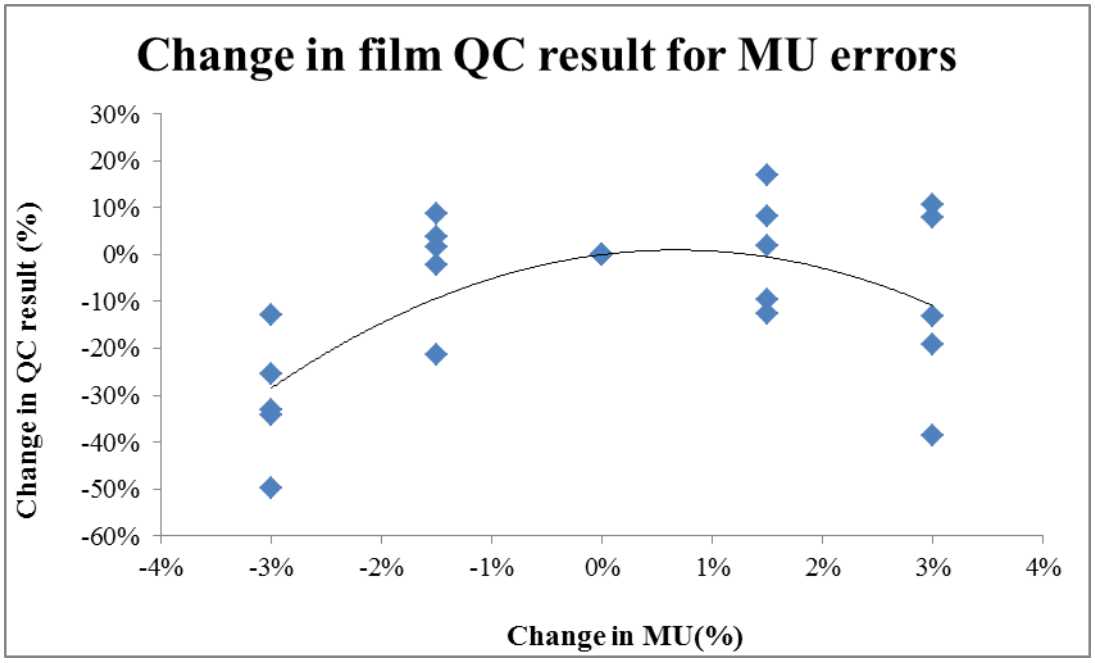

Figure 3.15: *Change in film QC results for systematic MU variations. Markers represent each individual QC measurement and the line is a 'guide to the eye'. With the applied γ-criterion of {2%;2mm}, smaller error magnitudes do not change the dose in the film plane enough to cause a systematic reduction in γ pass rate.*

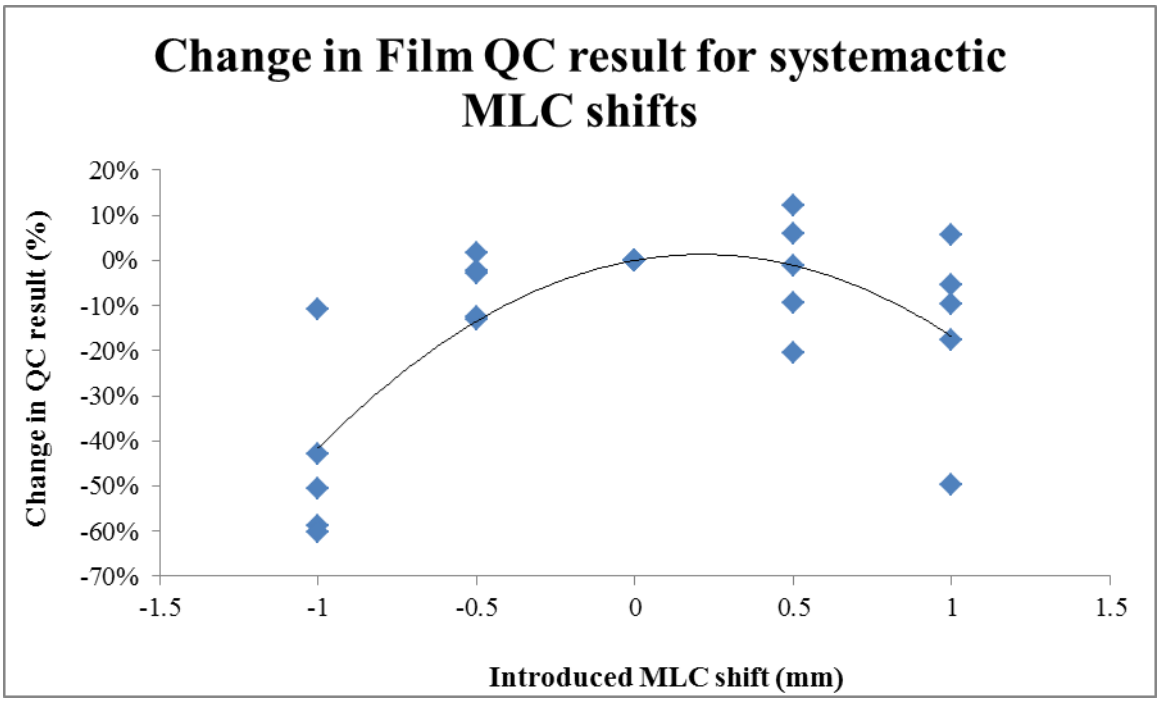

Figure 3.16: *Change in film QC results for systematic MLC shift errors. Markers represent each individual QC measurement and the line is a 'guide to the eye'. Again, smaller error magnitudes do not change the dose to individual pixels enough to cause a fail in the γ-criterion.*

3.4. ArcCheck Results Using Standard WBCC Set-up

Each error-free patient plan was measured using the ArcCheck technique on two separate occasions with the standard WBCC ArcCheck set up. The γ pass rates for both the {2%; 2mm} and {3%; 3mm} γ-criteria are given in **Table 3.23** for the clinical beam model and in **Table 3.24** for the adjusted beam model.

The measurements using the clinical beam model indicate that the dose map measured using the ArcCheck agreed with the TPS within the stated acceptance criteria for all measurements except one (patient 4, first measurement). However, the result for this plan was well within the acceptance criteria for the other measurement session. Based on these results, it appears that the ArcCheck accurately passes the majority of error-free plans. When using the adjusted beam model, only 3 out of 10 measurements were TPs.

Table 3.23: *ArcCheck γ pass rates for the error-free plans using configuration **A**. Both the passing rates for the 3%; 3mm ($P_\gamma^{\{3\%,3mm\}}$) and the 2%; 2mm ($P_\gamma^{\{2\%,2mm\}}$) gamma criteria are given. Acceptance criteria are 95% for the {3%; 3mm} γ-criterion and 85% for the {2%; 2mm} γ-criterion.*

Patient	Measurement 1		Measurement 2	
	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$
Patient 1	99.9%	98.8%	99.9%	99.8%
Patient 2	99.3%	94.1%	99.8%	97.1%
Patient 3	100.0%	97.7%	100.0%	99.6%
Patient 4	93.4%	79.0%	98.4%	88.2%
Patient 5	99.0%	94.4%	99.9%	97.0%

Table 3.24: *ArcCheck γ pass rates for the error-free plans using configuration **B**.*

Patient	Measurement 1		Measurement 2	
	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$
Patient 1	99.2%	92.1%	99.5%	91.4%
Patient 2	95.4%	79.0%	94.1%	74.8%
Patient 3	99.3%	86.3%	98.8%	84.7%
Patient 4	84.0%	57.6%	84.5%	56.4%
Patient 5	95.2%	80.3%	94.8%	77.7%

The γ pass rates are clearly lower for configuration **B** compared to configuration **A**. This large difference in results was further investigated by plotting the measured versus TPS dose maps for both beam models for patient 2 in **Figure 3.17**.

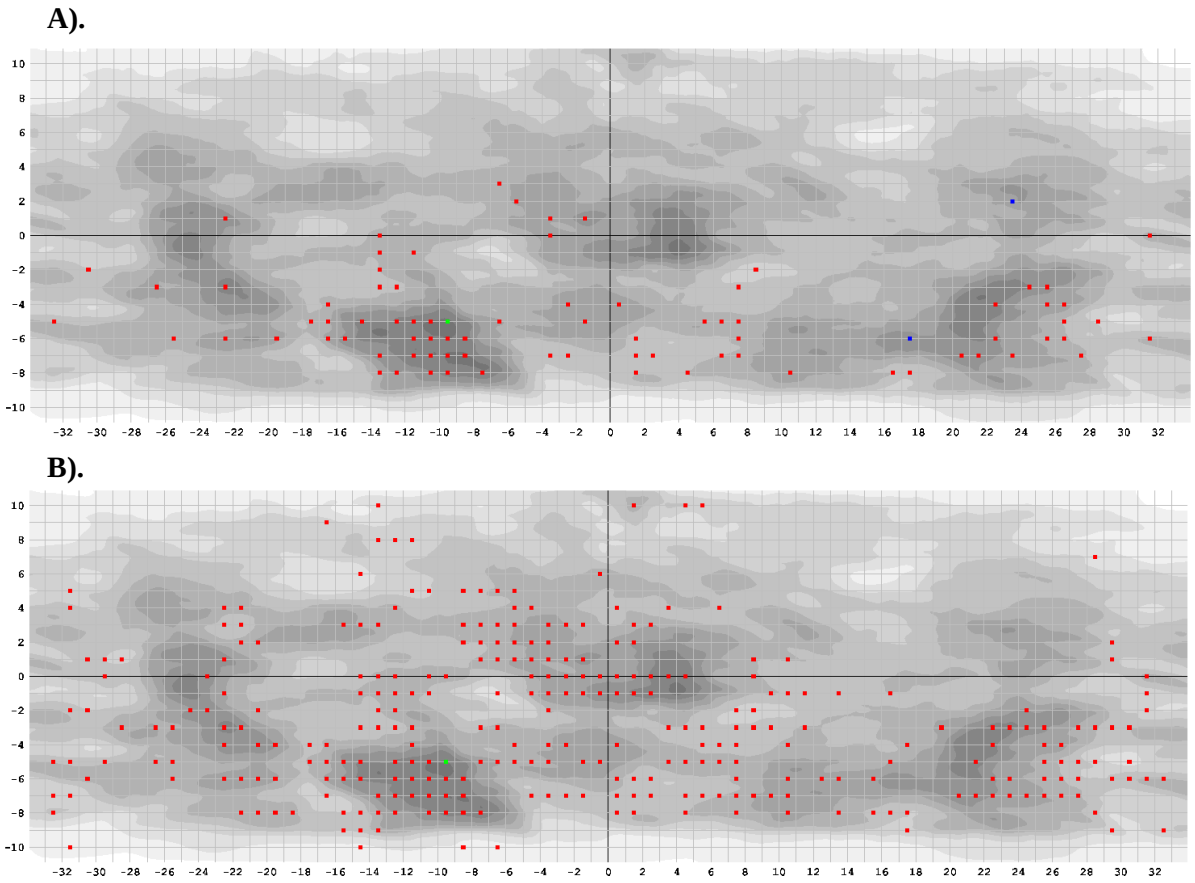

Figure 3.17: *Example ArcCheck dose difference map between measured data and TPS calculated data for the error-free plan of patient 2 using* ***(A)*** *the clinical beam model and* ***(B)*** *the adjusted beam model. The x and y axes represent distance in mm around the ArcCheck diameter and along the length of the ArcCheck respectively. The grey scale represents dose in Gy, ranging from 0Gy (white) to 1.5Gy (darkest grey). Red points represent diode locations where the ArcCheck measured dose was higher than calculated and exceeded the {2,; 2mm} γ-criterion, while blue points represent diodes where the measured dose is below calculated and the γ-criterion was exceeded.*

There are considerably more failure points for the adjusted beam model compared to the clinical beam model result (see also **Table 3.23** and **Table 3.24**). Considering that the same measured dose map is applied in both analyses, the difference in results could only be explained by a difference in the TPS calculated dose. This was confirmed by directly comparing the TPS calculated dose from both beam models (see **Figure 3.18**).

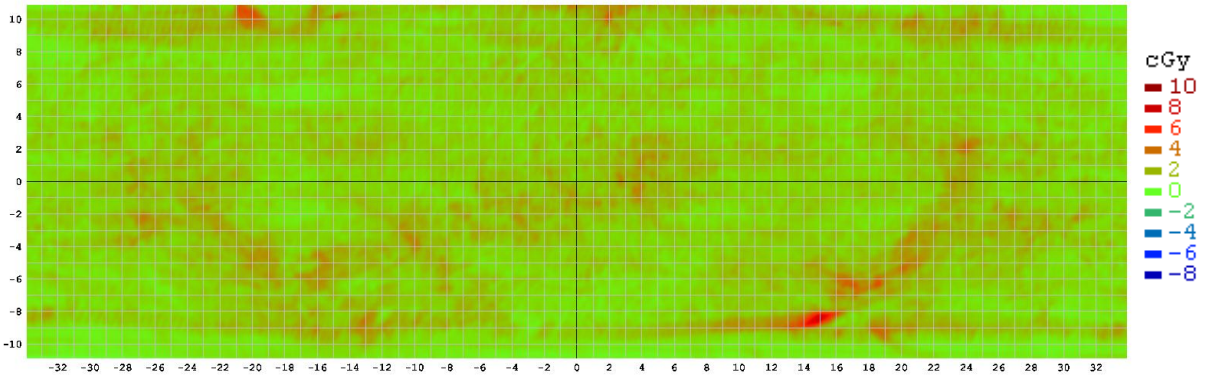

Figure 3.18: *Absolute dose difference map between verification plans calculated using the adjusted beam model and verification plan calculated using the clinical model for patient 2 (clinical model minus adjusted model). The colour scale indicates the magnitude of the absolute dose difference. In general, the map is green (no dose difference), but there are a large number of red areas, indicating the clinical beam model calculates a higher dose at these locations.*

As stated in section 3.1, the clinical beam model yielded 1.7 % higher TPS calculated doses on average for all verification plans compared to the adjusted beam model due to the change in DLG. While this was the case for all QC methods, this had the largest impact on QC results of the ArcCheck system.

ArcCheck QC measurements were carried out for all plans including introduced errors using the method outlined in section 2.6.4 over the course of one measurement session. These results for configuration **A** are given in **Table 3.25** which also indicates the γ pass rate for the {2%; 2mm} γ-criterion.

Table 3.25: *ArcCheck pass rate using the {2%; 2mm} γ-criterion ($P_\gamma^{\{2\%,2mm\}}$) and dichotomous classification for all plans containing introduced errors using configuration **A**.*

Plan error(s)	Patient 1		Patient 2		Patient 3		Patient 4		Patient 5	
	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome	$P_\gamma^{\{2\%,2mm\}}$ (%)	Outcome
MU 3 % low	**95.8%**	**FN**	**88.6%**	**FN**	**98.8%**	**FN**	**93.9%**	**FN**	**91.3%**	**FN**
MU 1.5% low	99.7%	TN	96.7%	TN	99.7%	TN	96.1%	TN	97.0%	TN
MU 1.5% high	91.4%	TN	**78.8%**	**FP**	90.8%	TN	**63.1%**	**FP**	**81.9%**	**FP**
MU 3% high	80.1%	TP	61.3%	TP	77.1%	TP	44.7%	TP	64.3%	TP
Output w gantry angle 8%	**95.7%**	**FN**	**88.9%**	**FN**	**97.4%**	**FN**	**91.8%**	**FN**	**86.9%**	**FN**
Output w gantry angle 4%	95.8%	TN	95.8%	TN	99.3%	TN	89.1%	TN	97.3%	TN
MLC closed 1 mm	**98.3%**	**FN**	**88.6%**	**FN**	**95.4%**	**FN**	**87.9%**	**FN**	**90.3%**	**FN**
MLC closed 0.5 mm	99.6%	TN	96.7%	TN	**99.6%**	**FN**	**91.2%**	**FN**	**97.3%**	**FN**
MLC open 0.5 mm	90.8%	TN	70.2%	TP	**91.1%**	**FN**	52.0%	TP	**74.5%**	**FP**
MLC open 1 mm	67.8%	TP	37.5%	TP	47.5%	TP	24.8%	TP	47.0%	TP
MLC translation 1mm	98.1%	TN	91.8%	TN	96.7%	TN	**78.3%**	**FP**	90.5%	TN
MLC SC 1 mm	**90.8%**	**FN**	81.2%	TP	92.9%	TN	62.9%	TP	85.6%	TN
MLC SC 2 mm	78.9%	TP	64.5%	TP	76.6%	TP	48.0%	TP	68.2%	TP
MLC BS 2 mm	**94.6%**	**FN**	74.8%	TP	**93.2%**	**FN**	60.6%	TP	82.4%	TP
MLC Chiasm 2 mm	-		**91.0%**	**FN**	97.8%	TN	78.6%	TP	93.8%	TN
MU 3% high, MLC 1mm closed	98.3%	TN	92.8%	TN	**98.6%**	**FN**	**85.3%**	**FN**	93.1%	TN
DLG 1.2 mm	94.0%	TN	**79.6%**	**FP**	**86.9%**	**FN**	**56.3%**	**FP**	**79.9%**	**FP**
ETSS X 1.5 mm	98.3%	TN	93.9%	TN	97.5%	TN	**79.3%**	**FP**	94.1%	TN

These results were then summarised into truth tables indicating (taking account of the clinical relevance of each error) the total number of TPs, TNs, FPs and FNs for both configuration **A** and **B** in **Table 3.26A - B**. The percentage of false results is higher for both configurations (36% for configuration **A**, 46% for configuration **B**) using the ArcCheck relative to both trPD and film results.

Table 3.26 A - B: *Truth tables for both beam models indicating the number of positive and negative results for ArcCheck measurements.*

Configuration **A**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	23	26
Not clinically relevant	10	40

Configuration **B**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	30	19
Not clinically relevant	27	23

3.4.1. S_1 and Sp_1

As for the earlier point dose and film results, plots of ArcCheck QC results against change in clinical relevance metrics were created (see **Figure 3.19A - B**).

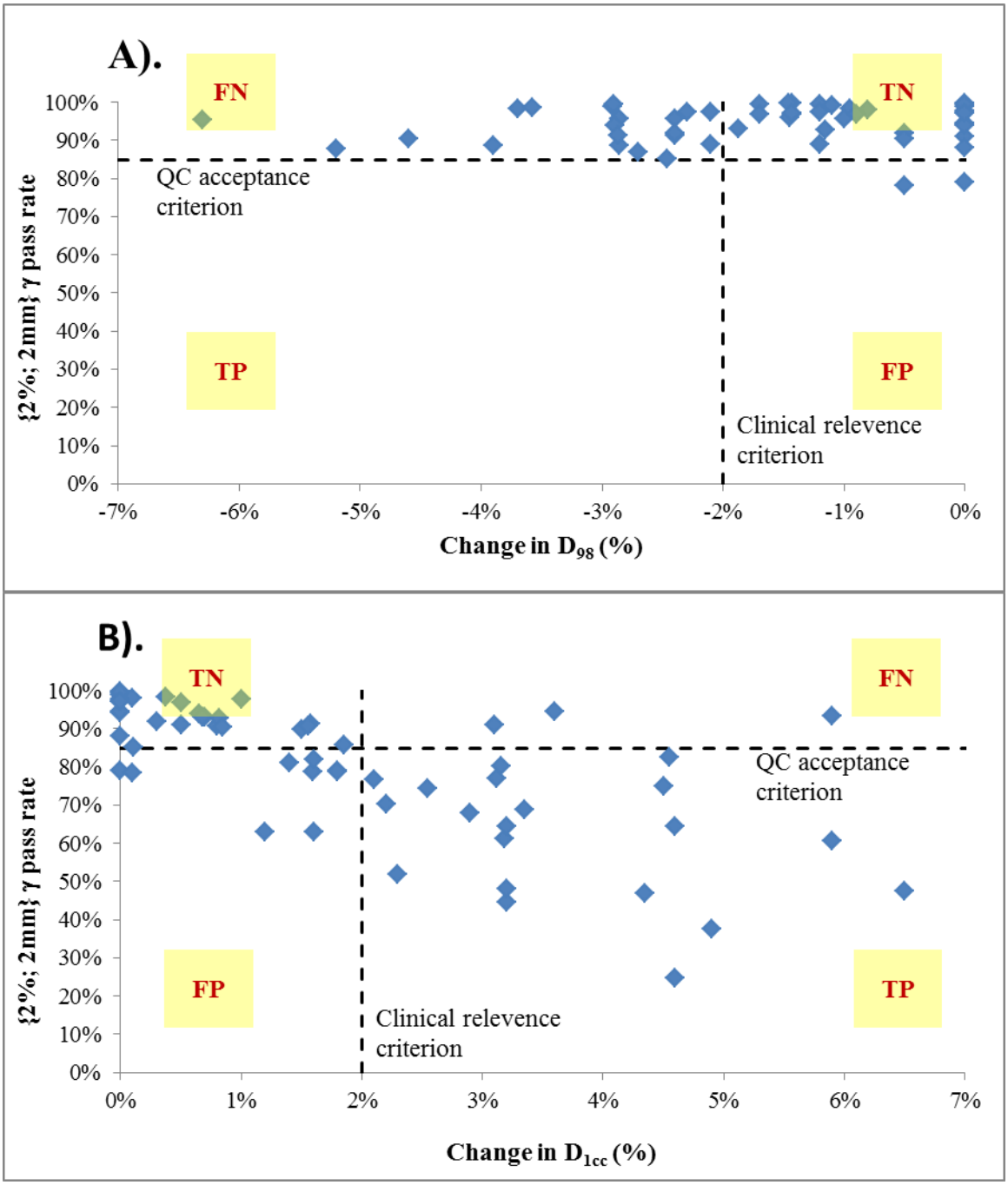

Figure 3.19 A (top) and B (bottom): *Plots of ArcCheck results against change in clinical relevance metrics with indications of QC passing criterion, clinical relevance criterion and regions corresponding to true and false positives and negatives.*

The results would be expected to follow a similar trend as for the film QC results (see **Figure 3.11A - B**) considering both methods employ γ analysis. The trends were similar when the change in D_{1cc} increased (**Figure 3.19B**). However, as the change in D_{98} decreased, the γ pass rates remained very high and no TP results were present. This is reflected in a high number of FNs and reduced sensitivity. S_1 and Sp_1 were calculated for all measurements using Equation 2.8 and Equation 2.12 and the results are given in **Table 3.27**.

Table 3.27: *S_1 and Sp_1 for the ArcCheck method using the current WBCC set-up as determined using the methods outlined in sections 2.7.1 and 2.8.1 for both configuration **A** and **B**.*

	Configuration A	**Configuration B**
S_1	46.9%	61.2%
Sp_1	80.0%	46.0%

Sp_1 for configuration **A** is comparable to the values obtained for the trPD and film methods but S_1 was considerably lower than the values for the other two methods. S_1 using configuration **B** was similar to the values obtained for the other methods, but Sp_1 was considerably lower than for the trPD and film methods.

ROC curves were created (**Figure 3.20**) for the ArcCheck technique for verification plans calculated using both configurations, with the optimal S_1 and Sp_1 values determined using the Youden index. These values, the AUC and optimal ArcCheck acceptance criterion for both configuration **C** and **D** are included in **Table 3.28**.

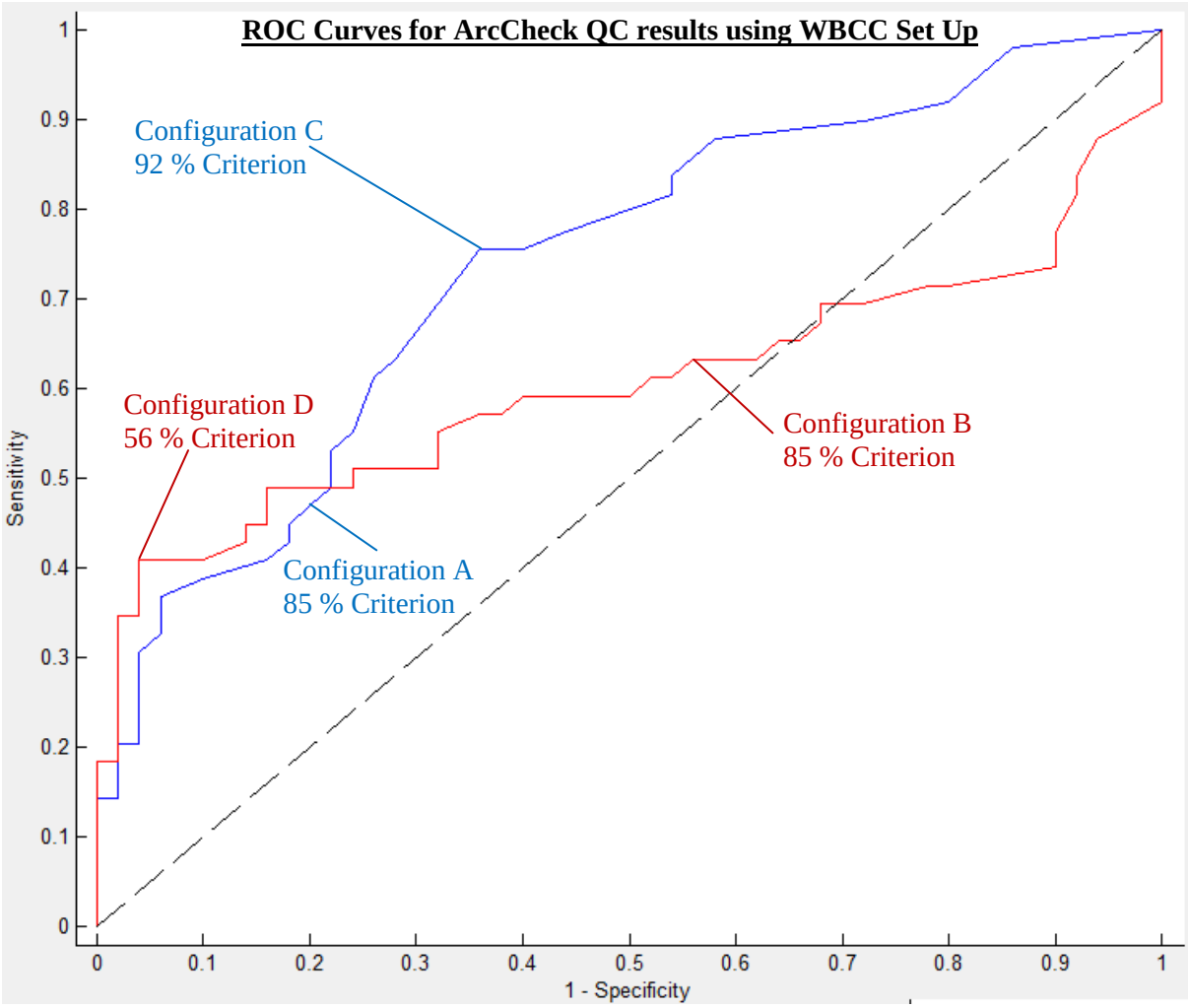

Figure 3.20: *ROC curves for ArcCheck verification measurements (using the WBCC set up) for both the clinical beam model (blue line) and the adjusted beam model (red line). The black dashed line represents the 0.5 area under the curve (AUC) value. The positions on the curve which correspond to configuration **A** and **B** (85% passing criterion) as well as the optimal acceptance criteria for each curve (configuration **C** and **D**, see **Table 3.28**) are also indicated.*

Table 3.28: *Metrics characterising the efficiency of the ArcCheck. S_1 and Sp_1 values are those determined using configuration **C** and **D**. The values in brackets represent the change in a given result from configuration **A** and **B** respectively.*

	Configuration C (change from A)	**Configuration D (change from B)**
AUC	0.74	0.60
S_1	75.5% (+28.6%)	40.8% (-20.4%)
Sp_1	64.0% (-16.0%)	96.0% (+50.0%)
Optimal QC acceptance criterion	92%	56%

As for the other methods, the efficiency of the ArcCheck may have been reduced due to an increased number of false results from OAR specific MLC shifts. Therefore, the ROC analysis was repeated without including any OAR specific MLC shift errors (**Figure 3.21**).

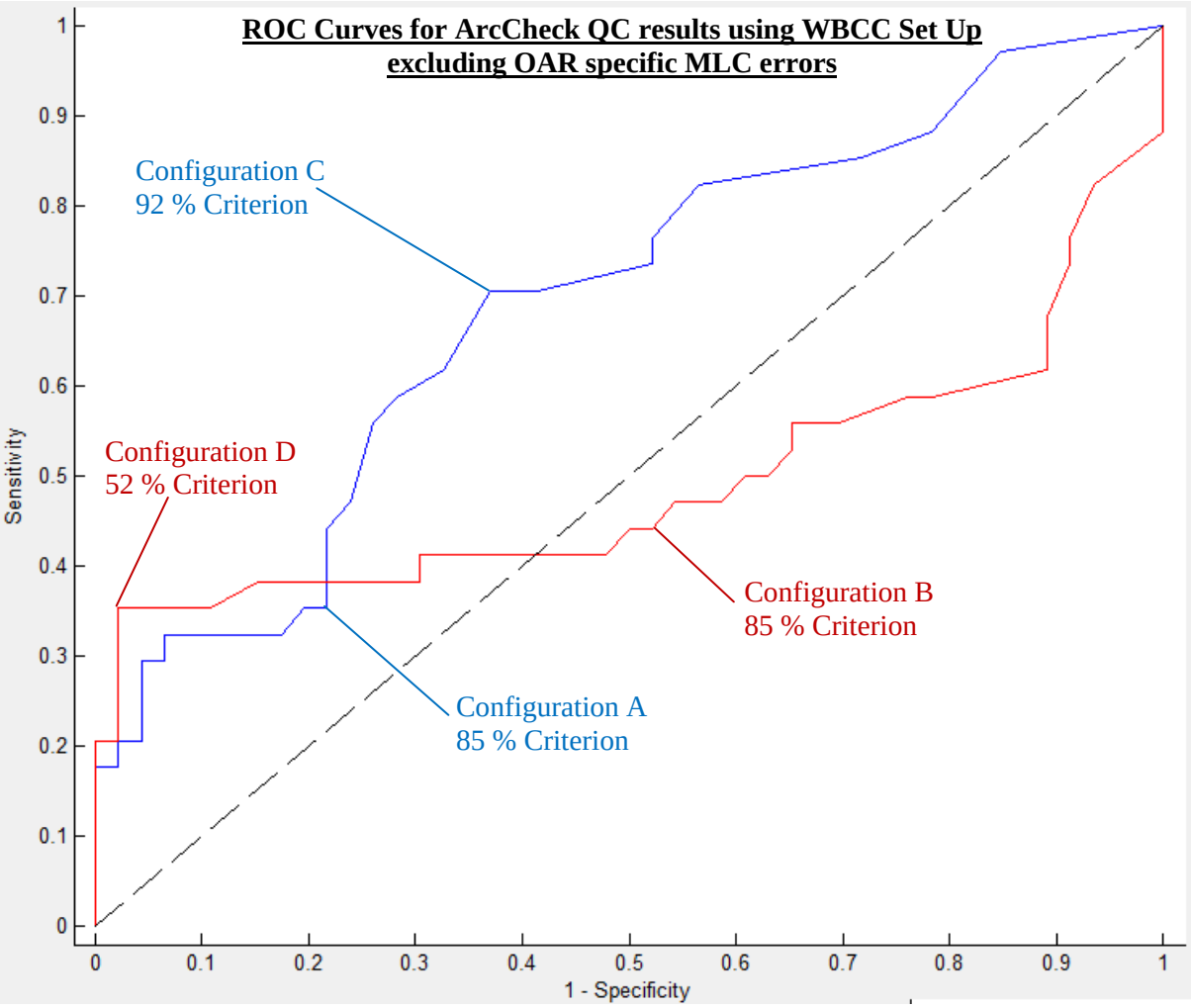

Figure 3.21: *ROC curves for ArcCheck results excluding OAR specific MLC errors (using the WBCC set up) for both the clinical beam model (blue line) and the adjusted beam model (red line).*

From these ROC curves, the AUC was calculated, and S_1 and Sp_1 were determined for the optimal QC passing criteria (see **Table 3.29**). The AUC for both configurations is virtually the same regardless of whether OAR specific errors were included or not. The exclusion of the OAR specific errors seemed to result in an improvement of either S_1 or Sp_1 depending on the applied configuration, but also yields a lower value for the complementary metric.

Table 3.29: *Metrics characterising the efficiency of the ArcCheck excluding OAR specific errors. S_1 and Sp_1 values are those determined using configuration **C** and **D**. The values in brackets represent the change in a given result from configuration **A** and **B** respectively.*

	Configuration C (change from A)	**Configuration D (change from B)**
AUC	0.69	0.60
S_1	70.6% (+28.6%)	35.3% (-20.4%)
Sp_1	63.0% (-17.0%)	97.8% (+51.8%)
Optimal QC acceptance criterion	92% (+7%)	52% (-33%)

3.4.2. S_2

S_2 was calculated as per Equation 2.10 for each plan containing introduced errors. The median sensitivity and range of sensitivities observed over all five patients are given in **Table 3.30**.

Table 3.30: *Median S_2 and range of S_2 over five patients for each plan containing introduced errors for both the {3%;3mm} and {2%;2mm} γ-criteria.*

Plan error(s)	**{3%;3mm} γ-criterion**				**{2%;2mm} γ-criterion**			
	Median S_2	**S_2 Range**			**Median S_2**	**S_2 Range**		
MU 3% low	N/A		-		146.3%	110.5%	-	526.3%
MU 1.5% low	N/A		-		N/A		-	
MU 1.5% high	N/A		-		N/A		-	
MU 3% high	50.0%	16.7%	-	100.0%	88.0%	72.8%	-	130.5%
Output w gantry angle 8%	158.6%	0.0%	-	166.7%	113.4%	78.0%	-	125.0%
Output w gantry angle 4%	N/A		-		N/A		-	
MLC 1 mm closed	135.2%	54.6%	-	189.1%	121.0%	49.5%	-	147.0%
MLC 0.5 mm closed	129.2%	100.0%	-	400.0%	178.6%	109.1%	-	256.3%
MLC 0.5 mm open	83.3%	50.0%	-	157.1%	109.4%	83.3%	-	120.7%
MLC 1 mm open	109.7%	90.2%	-	114.0%	93.1%	90.3%	-	101.5%
MLC 1 mm translation	100.0%	100.0%	-	100.0%	133.3%	100.0%	-	150.0%
MLC SC 1 mm	141.7%	100.0%	-	200.0%	113.3%	33.3%	-	142.9%
MLC SC 2 mm	95.4%	78.2%	-	110.3%	96.6%	84.3%	-	106.0%
MLC BS 2 mm	89.9%	0.0%	-	125.0%	94.2%	79.4%	-	105.4%
MLC Chiasm 2 mm	100.0%	100.0%	-	100.0%	75.0%	66.7%	-	100.0%
MU 3% high, MLC 1 mm closed	162.5%	0.0%	-	300.00%	139.4%	21.9%	-	162.5%

There are a number of entries in **Table 3.30** where N/A is entered instead of a value for S_2. This indicates that for that particular error mode and γ-criterion, the change in dose did not exceed the γ-

criterion and resulted in in 0% change in pass rate due to the introduction of the error for all patients (i.e. change in input is 0%, therefore S_2 cannot be determined).

3.4.3. S_3

S_3 was calculated as per Equation 2.11 for the {2%; 2mm} γ-criterion only using verification plans calculated using the clinical beam model. Errors that cause an increase in overall dose (MU increases or MLC open shifts) have much more negative values of S_3 compared to errors that cause a decrease in overall dose (MU decrease or MLC closed shifts). For these error modes, S_3 is positive for a number of cases (see **Table 3.31**), again indicating that the ArcCheck γ pass rates increase for a number of introduced errors.

Table 3.31: *Median S_3 and range of S_3 for each individual error type measured using the ArcCheck method implementing the {2%; 2mm} γ-criterion for plans calculated using the clinical beam model.*

Plan error	Unit	Median S_3	S_3 Range
MU 3% low	$\%.\%^{-1}$	-1.3	-2.8 - 1.9
MU 1.5% low	$\%.\%^{-1}$	0.0	-0.3 - 5.3
MU 1.5% high	$\%.\%^{-1}$	-8.3	-10.6 - -4.6
MU 3% high	$\%.\%^{-1}$	-10.0	-11.4 - -6.2
Output w gantry angle 8%	$\%.\%^{-1}$	-0.5	-1.3 - 0.5
Output w gantry angle 4%	$\%.\%^{-1}$	0.4	0.2 - 2.5
MLC 1 mm closed	$\%.mm^{-1}$	-2.3	-5.5 - 8.9
MLC 0.5 mm closed	$\%.mm^{-1}$	5.2	1.6 - 24.4
MLC 0.5 mm open	$\%.mm^{-1}$	-39.8	-54.0 - -13.2
MLC 1 mm open	$\%.mm^{-1}$	-50.2	-56.6 - -31.0
MLC 1 mm translation	$\%.mm^{-1}$	-1.0	-3.9 - -0.7
MLC SC 1 mm	$\%.mm^{-1}$	-8.8	-16.1 - -4.8
MLC SC 2 mm	$\%.mm^{-1}$	-12.9	-15.5 - -10.0
MLC BS 2 mm	$\%.mm^{-1}$	-6.0	-9.7 - -2.1
MLC Chiasm 2 mm	$\%.mm^{-1}$	-0.3	-1.6 - 0.1
DLG 1.2 mm	$\%.mm^{-1}$	-42.3	-79.8 - -14.5
ETSS X 1.5 mm	$\%.mm^{-1}$	-1.9	-5.9 - -1.0

Plots of error magnitude against change in γ pass rate are displayed for systematic MU errors and MLC shifts (**Figure 3.22** and **Figure 3.23**). For errors with a positive magnitude (either increasing output or opening the MLC), the γ pass rate dropped as the error magnitude was increased. However,

for errors with a negative magnitude (either decreasing the output or closing the MLC), the ArcCheck γ pass rate either stayed similar to that of the error free plan or increased slightly as the error magnitude was increased.

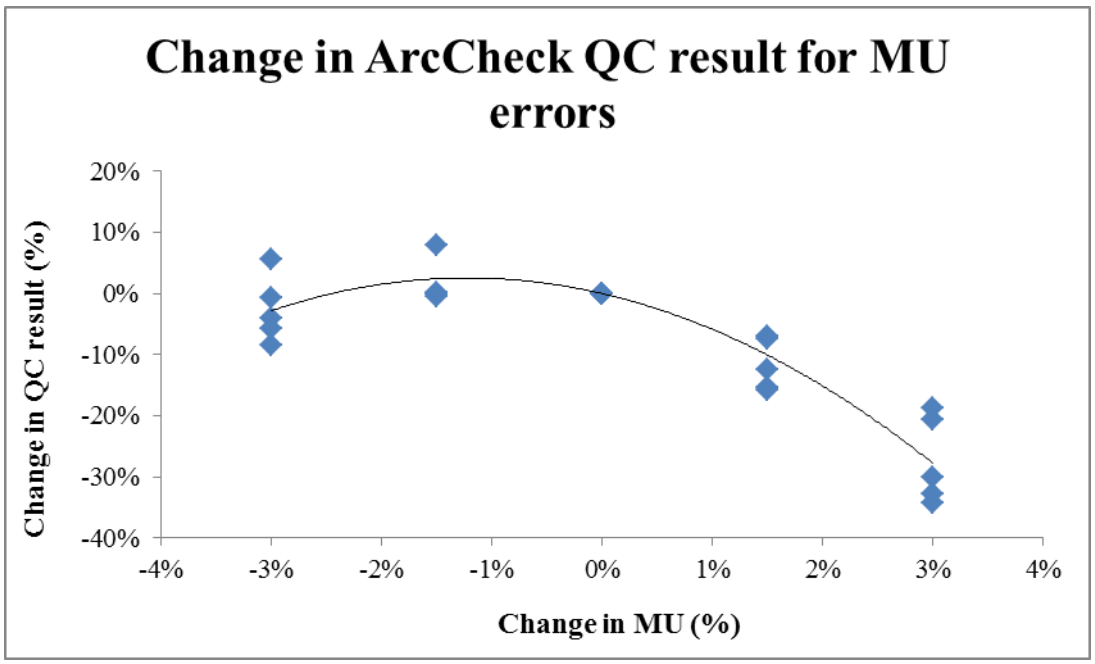

Figure 3.22: *Change in ArcCheck QC results for systematic MU errors. Markers represent each individual QC measurement and the line is a 'guide to the eye'. Note that the change in QC result is negative for increasing change in MU, but relatively constant for decreasing change in MU.*

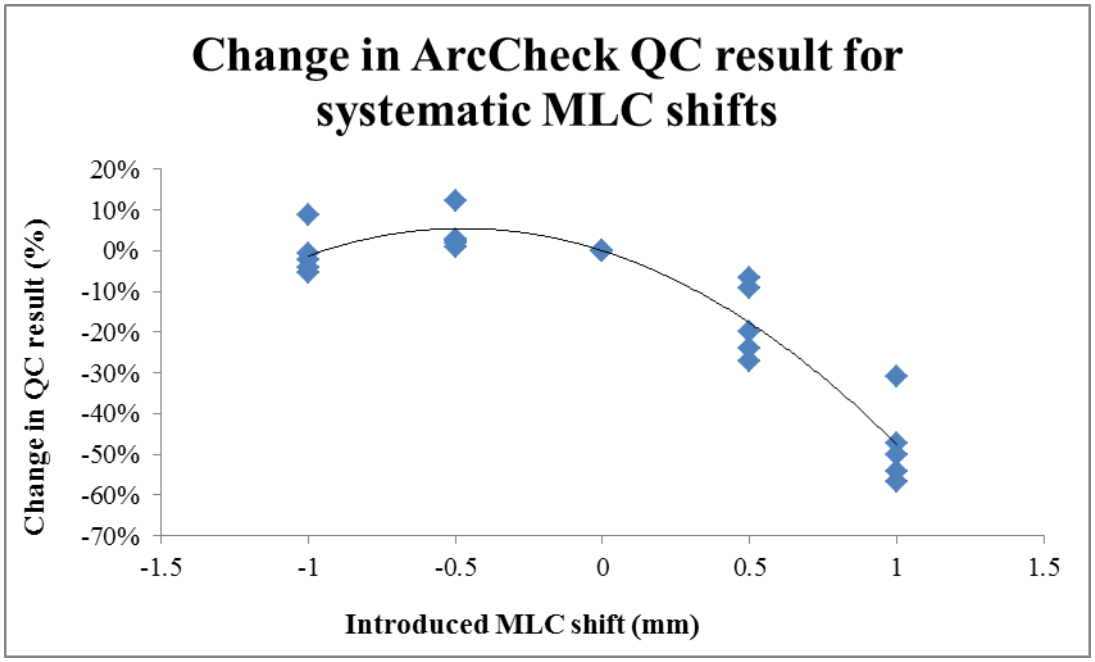

Figure 3.23: *Change in ArcCheck QC results for systematic MLC shift errors. Markers represent each individual QC measurement and the line is a 'guide to the eye'. Note that the change in QC result is negative for increasing (opening) MLC shift, but relatively constant for decreasing (closing) MLC shift.*

3.5. ArcCheck Results Using Recommended ArcCheck Set-up

All error-free plans were re measured using the ArcCheck method utilising the recommended ArcCheck set up outlined in section 2.6.4 with the results given in **Table 3.32** for configuration **A'**.

Table 3.32: *ArcCheck γ pass rates for the error-free plans that were determined using configuration **A'** Results for configuration **A** are given for comparison. Both the passing rates for the 3%; 3mm ($P_\gamma^{\{3\%,3mm\}}$) and the 2%; 2mm ($P_\gamma^{\{2\%,2mm\}}$) γ-criteria are given. Results are γ pass rates for comparison with verification plans calculated using the HUo and HC.*

Patient	Configuration A'		Configuration A	
	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$	$P_\gamma^{\{3\%,3mm\}}$	$P_\gamma^{\{2\%,2mm\}}$
Patient 1	99.60%	92.90%	99.9%	98.8%
Patient 2	95.70%	78.90%	99.3%	94.1%
Patient 3	95.70%	89.70%	100.0%	97.7%
Patient 4	87.00%	61.70%	93.4%	79.0%
Patient 5	94.60%	78.30%	99.0%	94.4%

These results were similar to those obtained earlier using configuration **B** and were worse than those obtained configuration **A** (see **Table 3.23** and **Table 3.24** in section 3.5.1).

All plans containing introduced errors were re-measured using the manufacturer's recommended set up. The results are summarised in **Table 3.33**.

Table 3.33: *ArcCheck pass rates using the {2%; 2mm} γ-criterion ($P_{\gamma}^{\{2\%,2mm\}}$) and dichotomous classification for all plans containing introduced errors.*

Plan error(s)	Patient 1		Patient 2		Patient 3		Patient 4		Patient 5	
	$P_{\gamma}^{\{2\%,2mm\}}$ (%)	Outcome	$P_{\gamma}^{\{2\%,2mm\}}$ (%)	Outcome	$P_{\gamma}^{\{2\%,2mm\}}$ (%)	Outcome	$P_{\gamma}^{\{2\%,2mm\}}$ (%)	Outcome	$P_{\gamma}^{\{2\%,2mm\}}$ (%)	Outcome
MU 3 % low	**99.9%**	**FN**	**98.8%**	**FN**	**99.9%**	**FN**	**90.7%**	**FN**	**98.7%**	**FN**
MU 1.5% low	99.2%	TN	93.4%	TN	97.8%	TN	**78.8%**	**FP**	92.4%	TN
MU 1.5% high	**76.7%**	**FP**	**57.0%**	**FP**	**72.8%**	**FP**	**43.6%**	**FP**	**56.9%**	**FP**
MU 3% high	57.4%	TP	38.4%	TP	54.8%	TP	29.5%	TP	41.0%	TP
Output w gantry angle 8%	**99.2%**	**FN**	**94.4%**	**FN**	**97.8%**	**FN**	**87.6%**	**FN**	**94.5%**	**FN**
Output w gantry angle 4%	98.5%	TN	90.4%	TN	95.7%	TN	**77.5%**	**FP**	90.5%	TN
MLC closed 1 mm	**99.3%**	**FN**	**95.2%**	**FN**	**98.6%**	**FN**	**92.3%**	**FN**	**95.4%**	**FN**
MLC closed 0.5 mm	99.1%	TN	94.5%	TN	**97.9%**	**FN**	**85.3%**	**FN**	**93.0%**	**FN**
MLC open 0.5 mm	**74.3%**	**FP**	43.9%	TP	57.7%	TP	32.9%	TP	**48.0%**	**FP**
MLC open 1 mm	41.7%	TP	20.5%	TP	25.4%	TP	14.4%	TP	26.6%	TP
MLC trans 1mm	92.3%	TN	**79.0%**	**FP**	86.3%	TN	**62.6%**	**FP**	**73.6%**	**FP**
MLC SC 1 mm	80.5%	TP	62.5%	TP	**75.1%**	**FP**	47.4%	TP	**61.5%**	**FP**
MLC SC 2 mm	64.4%	TP	49.2%	TP	55.7%	TP	36.5%	TP	49.6%	TP
MLC BS 2 mm	83.9%	TP	55.1%	TP	75.9%	TP	38.8%	TP	59.7%	TP
MLC Chiasm 2 mm	-		77.5%	TP	88.9%	TN	58.2%	TP	**75.2%**	**FP**
MU 3% high, MLC 1mm closed	94.7%	TN	86.8%	TN	**95.4%**	**FN**	78.3%	TP	**83.9%**	**FP**

These results are summarised in a truth table (**Table 3.34**), and take into account the clinical relevance of the errors introduced.

Table 3.34: *Truth table indicating the number of positive and negative results for ArcCheck measurements using the recommended ArcCheck set up based on the clinical relevance of an introduced error.*

Configuration **A'**	Measured Result	
	POSITIVE	NEGATIVE
Clinically Relevant	29	19
Not clinically relevant	19	17

From this truth table it was clear that utilising the manufacturer's recommended set up led to more false results (FNs and FPs) than utilising the WBCC set up (45% false results compared to 36%). The number of true results was reduced when using the recommended set up compared to the current WBCC set up (55% true results compared to 64%).

3.5.1. S_1 and Sp_1

ArcCheck results obtained using the HU override and with the heterogeneity correction applied were plotted against the clinical relevance metrics, ΔD_{98} and ΔD_{1cc} (**Figure 3.24A - B**). These plots were very similar to those observed for the ArcCheck results obtained using the WBCC configuration (see **Figure 3.19A - B**), in that there was a very low rate of TPs recorded for errors which result in a decrease in D_{98}. Furthermore, there are a number of results for which the change in D_{1cc} was less than 2%, but failed the ArcCheck acceptance criterion ({2%;2mm} γ pass rate < 85%). This increased the number of FP results and decreased the number of TNs and resulted in a low value for Sp_1.

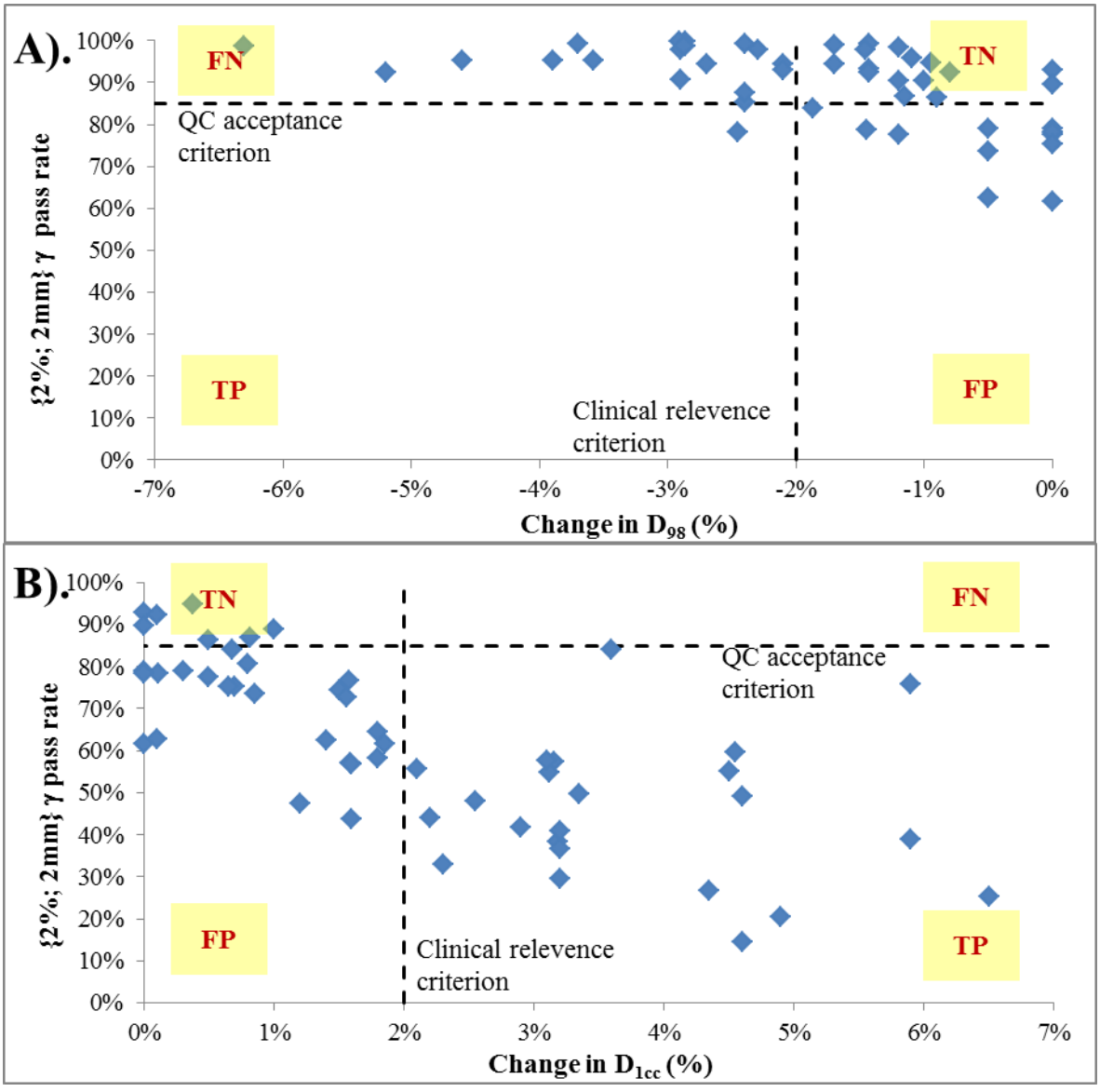

Figure 3.24 A (top) and B (bottom): *Plots of ArcCheck results against change in both clinical relevance metrics with indications of QC passing criterion, clinical relevance criterion and regions corresponding to true and false positives and negatives.*

S_1 and Sp_1 were then calculated for all measurements utilising the recommended ArcCheck set up using Equation 2.8 and Equation 2.12, and are displayed in **Table 3.35**.

Table 3.35: *S_1 and Sp_1 for the ArcCheck results obtained using configuration **A'**. Results for configuration **A** are given for comparison.*

	Configuration A'	**Configuration A**
S_1	60.4 %	46.9%
Sp_1	47.2 %	80.0%

The obtained values for S_1 and Sp_1 were very similar to those for the ArcCheck results using the WBCC set up and the adjusted beam model. S_1 was slightly higher for the recommended set up relative to the WBCC set up using configuration **A**, however Sp_1 was much lower.

The ROC curve (**Figure 3.25**) for plans measured using the HUo and HC was also generated, and the optimal S_1 and Sp_1 values were determined using the Youden index. These values are included in **Table 3.36** along with the AUC and optimal cut off threshold.

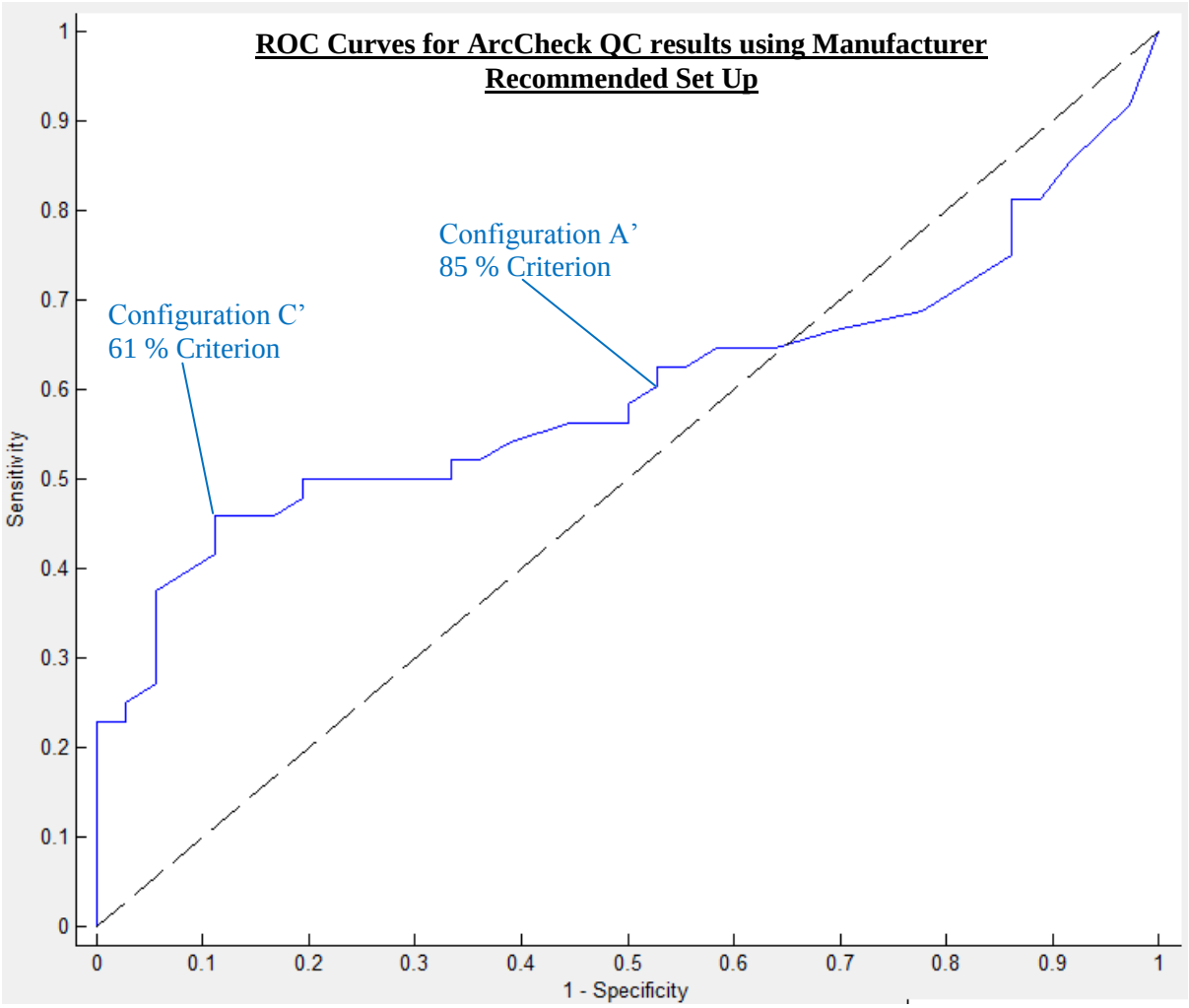

Figure 3.25: *ROC curve for ArcCheck verification measurements (using the manufacturer's recommended set up). The black dashed line represents the 0.5 area under the curve (AUC) value. The positions on the curve which correspond to configuration **A'** (85% passing criterion) as well as the optimal acceptance criterion (configuration **C'**, see **Table 3.36**) are also indicated.*

.

Table 3.36: *Metrics characterising the efficiency of the ArcCheck method using the recommended ArcCheck set up. The values in brackets represent the change in a given result from the current clinical QC acceptance criterion.*

	Configuration C' (change from A')
AUC	0.60
S_1	45.8 % (-14.6 %)
Sp_1	88.9 % (+41.7 %)
Optimal QC acceptance criterion	61 % (-24 %)

Changing the QC acceptance criterion could not achieve both a high S_1 and Sp_1 for a specific QC acceptance criterion. This is comparable to the results obtained for the ArcCheck method using the

WBCC set up. The ROC analysis was repeated, after removing OAR specific MLC shift errors (**Figure 3.26**).

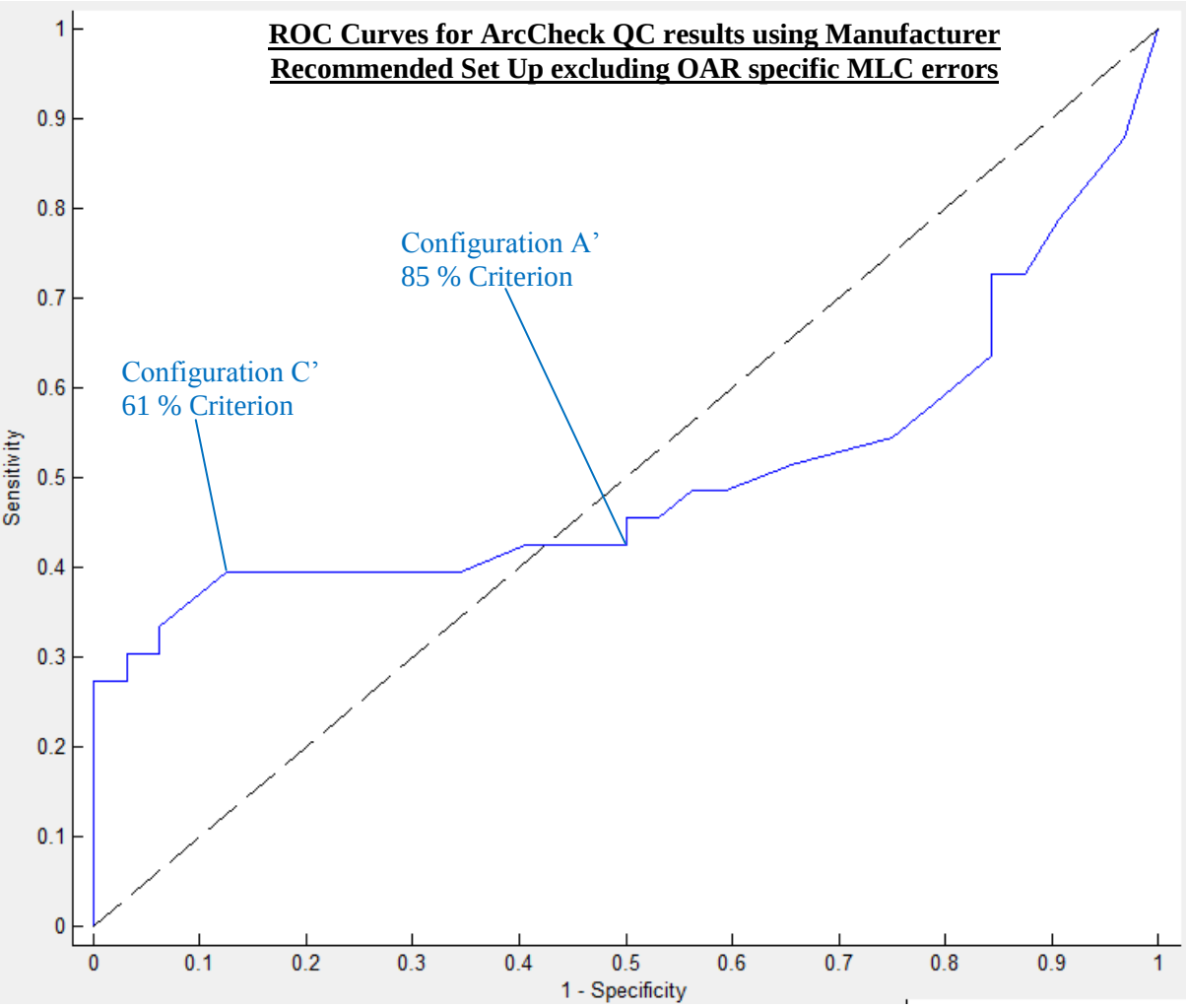

Figure 3.26: *ROC curve for ArcCheck verification measurements (using the manufacturer's recommended set up) excluding OAR specific MLC errors.*

From this ROC curve, the AUC, and S_1 and Sp_1 values for the optimal QC acceptance criteria were determined (**Table 3.37**). The AUC for the ArcCheck using the recommended configuration is 0.10 lower when the OAR specific MLC shift errors were excluded compared to these errors included in the ROC analysis.

Table 3.37: *Metrics characterising the efficiency of the ArcCheck method using the recommended ArcCheck set up excluding OAR specific MLC errors. The values in brackets represent the change in a given result from the current clinical QC acceptance criterion*

	Configuration C' (change from A')
AUC	0.50
S_1	39.8 % (-21.0 %)
Sp_1	87.5 % (+40.3 %)
Optimal QC acceptance criterion	61 % (-24 %)

3.5.2. S_2

S_2 was not determined using this ArcCheck configuration as time constraints did not allow this, and priority was given to completing the analysis of the other results.

3.5.3. S_3

S_3 was determined as per the methodology outlined in section 2.7.3 for measurements conducted with the HUo and HC correction applied. S_3 results were obtained for the {2%; 2mm} γ-criterion only. The results are given in **Table 3.38**. It can be seen that for every introduced error which led to a decrease in dose, the median S_3 value was more positive for the recommended set up relative to the WBCC set up, while for every error that caused an increase in dose the converse was true. This may indicate that there is an offset in either the ArcCheck measured or calculated dose and this will be further discussed in chapter 4.

Table 3.38: *Median S_3 and range of S_3 for each individual error type measured using the ArcCheck method implementing the {2%; 2mm} γ-criterion for plans calculated using the clinical beam model. The difference column is the ArcCheck median S_3 determined for measurements with the HUo and HC applied minus the corresponding S_3 value determined without either correction applied.*

Plan	Unit	Median S_3	S_3 Range			Difference
MU 3% low	$\%.\%^{-1}$	6.6	2.3	-	9.7	5.3
MU 1.5% low	$\%.\%^{-1}$	9.4	4.2	-	11.4	9.4
MU 1.5% high	$\%.\%^{-1}$	-12.1	-14.6	-	-10.8	-3.7
MU 3% high	$\%.\%^{-1}$	-11.8	-13.5	-	-10.7	-1.8
Output w gantry angle 8%	$\%.\%^{-1}$	1.9	0.8	-	3.2	1.4
Output w gantry angle 4%	$\%.\%^{-1}$	2.9	1.4	-	4.0	3.3
MLC 1 mm closed	$\%.mm^{-1}$	16.3	6.4	-	30.6	14.0
MLC 0.5 mm closed	$\%.mm^{-1}$	29.4	12.4	-	47.2	34.6
MLC 0.5 mm open	$\%.mm^{-1}$	-60.6	-70.0	-	-37.2	-20.8
MLC 1 mm open	$\%.mm^{-1}$	-51.7	-64.3	-	-47.3	-1.5
MLC 1 mm translation	$\%.mm^{-1}$	-0.6	-4.7	-	0.9	0.4
MLC SC 1 mm	$\%.mm^{-1}$	-14.6	-16.8	-	-12.4	-5.8
MLC SC 2 mm	$\%.mm^{-1}$	-14.4	-17.0	-	-12.6	-1.5
MLC BS 2 mm	$\%.mm^{-1}$	-9.3	-11.9	-	-4.5	-3.3
MLC Chiasm 2 mm	$\%.mm^{-1}$	-1.1	-1.8	-	-0.4	-0.9

4. Discussion

This chapter discusses the results and major findings of this study and is divided into the following sections:

- Section 4.1 – The importance of clinical relevance in the context of this study and the clinical relevance of the introduced errors.
- Section 4.2 – The results of the ROC analysis including:
  - The impact of varying the QC acceptance criteria.
  - The impact of varying the TPS beam model on patient-specific QC.
  - The influence of measuring at OAR locations compared to PTV locations.
  - The efficiency of the patient-specific QC methods at the WBCC compared to other studies in the literature.
- Section 4.3 – The results of the trPD and film QC methods utilising S_2 and S_3 are compared
- Section 4.4 – All ArcCheck results
- Section 4.5 – The limitations of each method
- Section 4.6 – The limitations of the ROC analysis method
- Section 4.7 – The advantages of being able to not only detect errors, but also resolve error modes
- Section 4.8 – The recommendations for the individual patient-specific QC methods and on patient-specific QC in general.
- Section 4.9 –Future work that should be carried out based on the outcomes of this study

4.1. Clinical Relevance of Errors

An important aspect of this study was the inclusion of the clinical relevance of the intentional errors that were introduced. A clinical relevance threshold was defined and different errors either just below or above this threshold were introduced to provide a meaningful sensitivity and specificity analysis.

Many intentional error studies in literature have investigated relatively large errors [45, 49, 83]. This usually yields extremely high sensitivities and specificities but this does not necessarily reflect the strong and weak points of the QC methods in a clinically more realistic scenario. By introducing intentional errors of smaller magnitudes it is possible to assess at which magnitude a given error mode becomes clinically relevant. The response of a given QC method to errors of these magnitudes will more accurately reflect the real-world efficiency of the QC method.

The clinical relevance of errors is not something that can be quantified in an absolute sense. Judgement on what is clinically relevant or not may vary between departments and different ROs within a department. The clinical relevance of an error will also vary between patients (biological response to radiation), and treatment site (dose prescription and fractionation used and proximity to OARs). This study looked to overcome some of these factors by focussing on a group of patients being treated for the same treatment site using the same prescription and fractionation. By using a single treatment indication, the aim was to provide more consistency about which errors were clinically relevant across all patients within the study.

The use of DVH metrics is a simple and straight-forward method to assess the clinical relevance of errors. This can either be done by looking at whether an error causes a DVH metric to drop below a certain threshold, or by looking at the change in a DVH metric caused by an error. Application of an absolute threshold causes the clinical relevance of an introduced error to be strongly dependent on the achieved plan quality as reflected by the DVH metrics of the error-free plan. By using the change in DVH metrics, the individual plan quality has less influence on clinical relevance compared to using absolute thresholds (see section 2.2). This is observed in the current study, as for most errors the clinical relevance of each error is the same over all patients. All ±3% MU errors, ±1 mm MLC open or closed shifts, 2 mm MLC open shifts near the SC and BS and 8% output decrease with gantry angle errors were clinically relevant for all patients. All ±1.5% MU errors, 1 mm MLC translation shifts and

4% output decrease with gantry angle errors were not clinically relevant for any patients. The only delivery errors for which the clinical relevance was plan dependant were ±0.5 mm MLC shift errors, 2 mm MLC open shift errors near the chiasm and 1 mm MLC open shifts near the SC. This was similar for TPS modelling errors as well, where the DLG modelling error was clinically relevant for only one out of five patients and the ETSS error was not clinically relevant for any patient. However, these beam modelling parameters induce an overall departmental systematic error. The DLG change of 0.8 mm resulted in a median ΔD_{98} of -1.7 %. This is below our 2 % criteria for clinical relevance but induces a considerable systematic reduction in dose coverage that can make other small errors clinically relevant that could be ignored otherwise. Although the introduced ETSS error was not clinically relevant for any patient in this study, a separate departmental investigation showed that this type of error could be clinically relevant for other types of treatment techniques and indications with a larger contribution of small fields such as stereotactic treatments [84]. In particular for small field sizes, the ETSS values have a larger effect on partial source occlusion [62]. These considerations emphasize the importance of investigating and minimising systematic errors on a departmental level.

While the criteria used in this study provide a logical determination of clinical relevance, they are still dependent on the optimisation of the clinical treatment plans and therefore there is still likely to be large variations between departments based on different methods for optimising treatment plans. This limits the applicability of the current study results for other departments. One method to overcome this may be to link the clinical relevance of any error to a decrease in TCP and/or an increase in NTCP. However, TCP/NTCP calculations have a large inherent uncertainty and could potentially impact the results of this study considerably due to the limited number of patients. This would only further limit the applicability and would require extra time. With the limitation on the time available for this study, it was therefore decided to prioritise other aspects of this study and not include TCP/NTCP calculations.

Another aspect is how the clinical relevance of an error varies with plan complexity. Further research into this aspect might be an interesting extension of this study. The clinical relevance of errors may correlate with different metrics used to measure plan complexity. For example, a 1 mm systematic MLC open shift may have a larger impact on a plan with high average MLC leaf travel or small average leaf pair opening (ALPO) compared to a plan with low average MLC leaf movement or large ALPO.

4.2. Optimising the Efficiency of WBCC QC Methods

For efficient patient-specific QC, it is important to optimise the number of TPs and TNs, and to limit the number of FPs and FNs. The metrics S_1, Sp_1 and the AUC are relevant when comparing the efficiency of the various QC methods, or when optimising the efficiency using the Youden index [82]. A summary of these metrics for the various configurations investigated in this study is therefore included in **Table 4.1** together with the results from comparable studies in literature. The classification of whether a QC method can be considered to be effective or not may depend on the application and the associated risks. For instance, the minimum required values for S_1 and Sp_1 may be different for VMAT and conventional fractionation schemes compared to stereotactic treatments applying a limited number of treatment fractions when the impact of not detecting certain errors is much larger. A general set of guidelines for ROC AUC analysis was proposed by Greiner et al. [79], who suggested five separate categories for AUC values:

$AUC = 0.5$	non-informative test
$0.5 < AUC \leq 0.7$	less accurate test
$0.7 < AUC \leq 0.9$	moderately accurate test
$0.9 < AUC < 1.0$	highly accurate test
$AUC = 1.0$	perfect test

An important aspect to consider when discussing our results is the uncertainty of the values obtained for the various metrics. McKenzie et al. investigated various QC methods and assessed the corresponding AUC of each method for a range of acceptance criteria as well as the 95% confidence

interval (CI) using a bootstrapping procedure [85]. They obtained CIs as large as 0.38 - 0.9 for the MapCheck device with an AUC of 0.65 for one specific configuration. No 95% C.I.s are (yet) available for the metrics in this study due to time constraints. However, in the light of the uncertainties reported by McKenzie et al., we estimated that the uncertainty in the observed values for S_1 and Sp_1 will be at least 10% and included this uncertainty estimate in our analysis.

For the clinically applied configuration **A**, the efficiency of trPD and film measurements for the PTV locations at the WBCC are comparable to those reported in literature and falls under the moderately accurate test category. Both trPD and film measurements at PTV locations showed higher Sp_1 values (range 80 – 91 %) compared to the S_1 values (63 – 65 %) which indicates that our current patient-specific QC is good at passing plans without errors, but is not particularly good at failing plans that contain errors. In contrast, the efficiency of trPD and film measurements for OAR locations at the WBCC is low. For trPD measurements, the poor efficiency for OAR locations is likely caused by the high dose gradient at these locations considering that the distance to agreement (DTA) was less than 2 mm for all OAR location measurements. This indicates that the efficiency of trPD measurements at OAR locations could be improved by increasing the accuracy with which the phantom and detector can be positioned. This can be achieved by using a cone-beam CT matching procedure to position the plastic water slab phantom prior to trPD QC measurements.

For EBT film measurements at the OAR plane using configuration **A**, the AUC is as low as 0.50 which is equivalent to completely random test results. The finite accuracy of film positioning in the direction perpendicular to the film plane combined with a high dose gradient in that direction may contribute to the low efficiency for the OAR locations as well but there is at least one other cause that may be related to the low efficiency for these measurements: In most of the cases, the dose distribution at the OAR film planes also includes part of the PTV and the current implementation of the γ-analysis focusses on the correctness of the dose delivered to the PTV. As the dose gradient

perpendicular to the film plane at these PTV locations is generally not higher than those at the film planes including only the PTV, the dose gradient at the OAR planes cannot explain the poor efficiency of film verification measurements for OAR planes. The γ-analysis currently applied in the film dosimetry software uses global dose deviations (dose deviations relative to the average PTV dose) *and* excludes points in the film plane with a dose below 50% of the average PTV dose. This configuration of the γ-analysis is not uncommon, but focusses on the correctness of the dose delivered to the PTV and largely ignores delivery errors that affect OARs. It is possible that the change in dose to an OAR is recorded by the film, but that this region is excluded from the γ-analysis, which is likely the main reason for the low efficiency of film dosimetry at the OAR locations. Modifying the film analysis software to allow γ-analysis on an individual OAR structure would likely improve the efficiency of film measurements made at OAR planes. However, it must be noted that improvements of the TPS beam model as for configuration **B** had a large impact on the efficiency of film verification measurements at the OAR locations (see further below).

The efficiency of the ArcCheck measurements at the WBCC was generally low and did not improve noticeably upon changing the configuration (including changing the beam model, acceptance criteria or software corrections). Analysis of the ArcCheck results focussing on the sensitivity metrics S_2 and S_3 highlighted a potential problem with this device. For that reason, most of the discussion on the ArcCheck results will be discussed separately in section 4.4.

Table 4.1: *Comparison of sensitivity, specificity and AUC results of various studies*

Study	QC method	Acceptance criteria	S_1	Sp_1	AUC	Error types	Golden standard	Notes
McKenzie et al. [85]	cc04 ion chamber	±3% dose deviation	47%	100%	0.94	Errors detected by golden standard	Ion chamber array	
	ArcCheck	Pγ > 90%; {3%;3mm}	60%	89%	0.84			
	EDR2 Film	Pγ > 90%; {3%;3mm}	60%	89%	0.84			
Kry et al. [47]	Ion Chamber	±3% dose deviation	25%	90%	0.66	Errors detected by golden standard	IROC phantom	
	EDR2 Film	Pγ > 90%; {3%;3mm}	33%	82%	0.70			
	MapCheck	Pγ > 90%; {3%;3mm}	14%	94%	0.61			
Kim et al. [44]	EBT2 Film	Pγ > 90%; {2%;2mm}			0.78 - 0.82	Intentional 0.5 mm MLC open and close shifts	Presence of error	SBRT
	MapCheck				0.72 - 0.88			
Bojechko et al. [45]	(in vivo) portal dosimetry	{3%; 3mm}			0.70-0.77	Intentional ±3% MUs	Presence of error	
					0.55-0.63	Intentional ±1 mm systematic MLC shift		
Aristophanous et al. [86]	ArcCheck	Pγ > 90%; {3%;3mm}	50%	89%		Detected errors of golden standard	Ion Chamber on central axis	
Coleman et al. [87]	ArcCheck	Pγ > 90%; {2%;2mm}	82%			Intentional MLC errors causing 3% and 5% DVH dose changes	Presence of error	
Fredh *et al.* [49]	Delta4	Pγ > 95%; {2%;2mm}	75%	100%		Intentional errors: +3% MU, systematic 2mm and 4mm MLC open shifts, and 2° and 5° collimator rotation	Presence of error	VMAT for 2 Brain, 1 H&N, and 1 prostate treatment
	Octavius		40%	100%				
	COMPASS		40%	100%				
	EpiQA		100%	75%				
This Study Configuration **A**	trPD	±2% dose deviation	65%	91%	0.79	Various MU, MLC and TPS modelling errors	Clinical relevance	PTV locations
	trPD	±2% dose deviation	77%†	90%†	0.86†			PTV locations†
	trPD	±2% dose deviation	80%	26%	0.65			OAR locations
	EBT film	Pγ > 85%; {2%;2mm}	61%	83%	0.77			PTV locations
	EBT film	Pγ > 85%; {2%;2mm}	74%†	81%†	0.83†			PTV locations†
	EBT film	Pγ > 85%; {2%;2mm}	33%	63%	0.50			OAR locations
	ArcCheck	Pγ > 85%; {2%;2mm}	47%	80%	0.74			All locations
	ArcCheck	Pγ > 85%; {2%;2mm}	35%†	78%†	0.69†			All locations†

† *Values obtained after excluding OAR specific intentional errors*

Table 4.1: *Continued*

Study	QC method	Acceptance criteria	S_1	Sp_1	AUC	Intentional errors	Gold standard	Notes
This Study Configuration **B**	trPD	±2% dose deviation	63%	80%	0.79	Various MU, MLC and TPS modelling errors	Clinical relevance	PTV locations
	trPD	±2% dose deviation	68%†	82%†	0.83†			PTV locations†
	trPD	±2% dose deviation	80%	65%	0.70			OAR locations
	EBT film	Pγ > 85%; {2%;2mm}	65%	81%	0.80			PTV locations
	EBT film	Pγ > 85%; {2%;2mm}	62%†	79%†	0.78†			PTV locations†
	EBT film	Pγ > 85%; {2%;2mm}	60%	90%	0.91			OAR locations
	ArcCheck	Pγ > 85%; {2%;2mm}	61%	46%	0.60			All locations
	ArcCheck	Pγ > 85%; {2%;2mm}	44%†	48%†	0.60†			All locations†
This Study Configuration **C**	trPD	±**1.9%** dose deviation	67%	89%	0.79	Various MU, MLC and TPS modelling errors	Clinical relevance	PTV locations
	trPD	±**1.9%** dose deviation	79%†	88%†	0.86†			PTV locations†
	trPD	±**4.0%** dose deviation	73%	78%	0.65			OAR locations
	EBT film	Pγ > **87%**; {2%;2mm}	74%	79%	0.77			PTV locations
	EBT film	Pγ > **87%**; {2%;2mm}	82%†	77%†	0.83†			PTV locations†
	EBT film	Pγ > **93%**; {2%;2mm}	73%	42%	0.50			OAR locations
	ArcCheck	Pγ > **92%**; {2%;2mm}	76%	64%	0.74			All locations
	ArcCheck	Pγ > **92%**; {2%;2mm}	71%†	63%†	0.69†			All locations†
This Study Configuration **D**	trPD	±**1.5%** dose deviation	74%	73%	0.79	Various MU, MLC and TPS modelling errors	Clinical relevance	PTV locations
	trPD	±**1.5%** dose deviation	79%†	75%†	0.83†			PTV locations†
	trPD	±**4.5%** dose deviation	73%	74%	0.70			OAR locations
	EBT film	Pγ > **92%**; {2%;2mm}	80%	71%	0.80			PTV locations
	EBT film	Pγ > **88%**; {2%;2mm}	62%†	88%†	0.78†			PTV locations†
	EBT film	Pγ > **88%**; {2%;2mm}	87%	90%	0.91			OAR locations
	ArcCheck	Pγ > **92%**; {2%;2mm}	41%	96%	0.60			All locations
	ArcCheck	Pγ > **92%**; {2%;2mm}	35%†	98%†	0.60†			All locations†

† *Values obtained after excluding OAR specific intentional errors*

4.2.1. Impact of Optimising Acceptance Criteria on the Efficiency of Patient-Specific QC

Optimisation of the efficiency of the QC methods based on the Youden index only selects a different point on the ROC curve and doesn't change the AUC for a given ROC curve. The ROC curves in this study were generated by changing the QC-acceptance criteria, so when the threshold for accepting dose deviations is adjusted, for instance in the comparison between configuration **A** and **C**, the AUC does not change. However, each ROC curve applies for a film QC method for a specific γ-criterion and a change of this parameter (e.g., {3%; 3mm} instead of {2%; 2mm}) *would* change the ROC curve and the AUC. The latter adjustment of the acceptance criteria is not included in the discussion in this paragraph. For a number of QC methods using either TPS beam model, the net improvement in efficiency by adjusting the acceptance criteria is less than 10%, which is within the estimated uncertainty of the metrics and is therefore questionable whether this adjustment should be implemented at all. For other QC methods, an improvement in sensitivity is made at the expense of the specificity or vice versa. For these cases, the implementation of this adjustment should be considered but also requires careful consideration whether a reduction in FNs at the cost of a higher number of FPs is justifiable. In addition, it is also possible that other factors such as the positioning accuracy for the trPD method and the inadequate film analysis for OAR planes are reducing the efficiency of these QC methods and changing the acceptance criteria might compensate for these factors. Only for trPD measurements at the OAR locations did changing the acceptance criteria for dose deviations from 2% to 4% result in a considerable improvement (52%) in Sp_1 without decreasing S_1. However, this low Sp_1 is likely to be due to the reduced positional accuracy in the high dose gradients at these OAR locations. Therefore, it is preferable to initially attempt to improve the efficiency of trPD measurements by improving the accuracy of the detector positioning rather than by adjusting the acceptance criterion.

The improvement in efficiency for film measurements at the OAR locations by changing the threshold from 85% to 88% minimum pass rate resulted in a 27% increase of S_1 while Sp_1 was unchanged.

Although this adjustment does result in a large increase in efficiency, it does not resolve the fundamental problem with the current EBT film analysis for OAR film planes described above.

4.2.2. Impact of Adjusting the TPS Beam Model on the Efficiency of Patient-Specific QC

The adjustment of the TPS beam model did improve the calculation accuracy of the beam penumbra considerably (see appendix 3.A). By comparing the metrics in **Table 4.1** for configuration **B** with respect to those for configuration **A**, it can be seen that adjustment of the TPS beam model does not provide an uniform improvement for all QC methods at all locations. A large improvement is obtained for film dosimetry at the OAR plane resulting in an increase of the AUC from 0.50 to 0.91 when the adjusted TPS beam model is used (Configuration **B**) instead of the clinically applied beam model (Configuration **A**) while both S_1 and Sp_1 increase by 27%. For the trPD measurements at the OAR locations, the use of the adjusted TPS beam model (configuration **B**) results in an increase of Sp_1 by 39% while S_1 and the AUC remain practically unchanged, but for trPD and film measurements at the PTV locations, a decrease in S_1 and Sp_1 ranging between -8 and -14% is observed for the majority of cases. The improvement in efficiency for the OAR locations is considerably larger than the estimated uncertainty of the metrics of approximately 10%. This phenomenon is therefore considered to be real and can be tentatively explained by the increased modulation of the dose delivery through more intense MLC shielding around OARs which is more accurately described by the adjusted TPS beam model. However, more investigation is required to confirm this explanation.

Varying beam model parameters may directly affect the response of a QC system in the form of a systematic offset between measured and calculated data. This is possible since some beam modelling parameters do not correspond to physical characteristics of the treatment machine, but are in a sense "fudge factors" to allow a better fit of measured and calculated data. Both the DLG and ETSS parameters can be considered to fall into this category. After these parameters have been optimised to minimise the deviations between TPS calculations and measurement of the beam penumbra, the

corresponding DLG may not be desirable for clinical applications if the residual (smaller) deviations constructively interfere and cause a net residual deviation during plan delivery [84]. In this study, as calculated by the TPS, the median change in D_{98} from using the two different DLGs was -1.7 %. One could hypothesise that it is reasonable to conclude that as this difference is less than 2.0 %, the difference between DLG values is negligible and either value could be used for VMAT planning purposes. This hypothesis is reinforced by the point dose results which showed the median integral dose difference for the plans with the 2.0 mm DLG was -0.6 %, while for the 1.2 mm DLG it was +0.8 %. Again, these values both agree with each other within 2.0 %, and are both within 1.0 % of zero. However, even a 1.0 % offset in measurement could have an effect on the QC result when an additional error is applied. As an example, a subset of the point dose results for patient 4 is included in **Table 4.2**.

Table 4.2: *Subset of integral point dose results for patient 4 for two different DLG values. ΔPD represents the change in integral point dose deviation relative to the original plan. FN results are highlighted in orange.*

Plan	DLG 2.0 mm			DLG 1.2 mm		
	Integral PD deviation (%)	**ΔPD (%)**	**Result**	**Integral PD deviation (%)**	**ΔPD (%)**	**Result**
Original Plan	-1.6	-	TN	0.2	-	TN
MU 3% increase	1.3	+2.9	FN	3.2	+3.0	TP
MU 3% decrease	-4.9	-3.6	TP	-2.7	-2.9	TP

This example shows how the result of the QC test is influenced by the choice of beam model for an intentional error, while both error-free plans give the same QC result (TN). For the 3% MU increase error a difference in QC result is observed, even though the change in integral deviation is very similar for both DLG models. So even though different DLG values may lead to acceptable plans, the value of the DLG may have a considerable impact on the sensitivity of QC methods.

Overall, the different beam models led to large differences in QC results as can be seen from plots of the QC results against clinical relevance for each QC method using different beam models (see **Figure 4.1**).

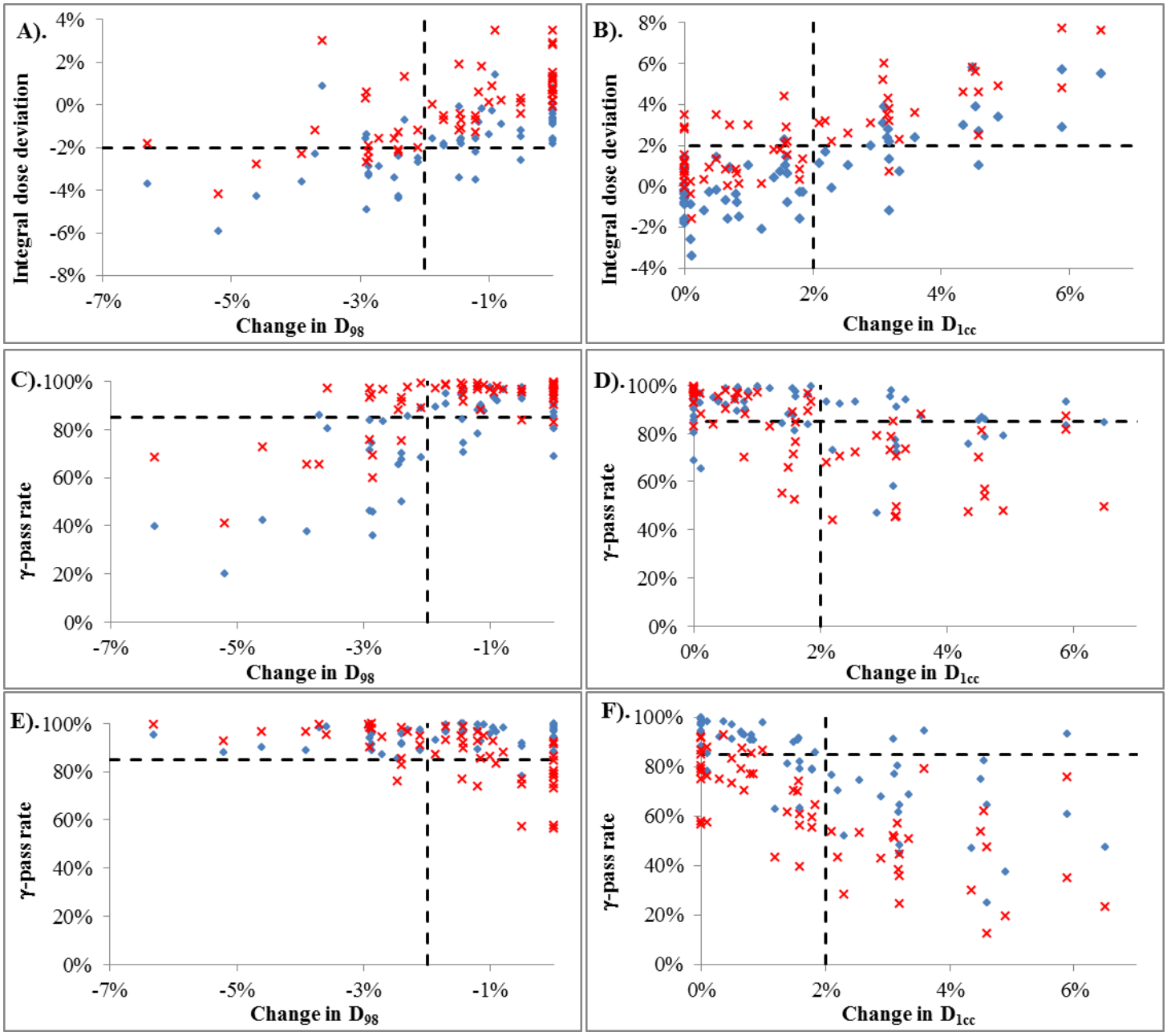

Figure 4.1 A-F: *Plots of QC results against change in clinical relevance metrics. Plots A and B are for trPD results. Plots C and D are for film results and plots E and F are for ArcCheck results. Blue markers represent QC results obtained using the clinical beam model; while red crosses represent QC results obtained using the adjusted beam model.*

Starting with the trPD plots (**Figure 4.1A - B**), that the most prominent effect of changing beam models on each QC measurement is the vertical systematic shift of 1.7 %. This is observed as a systematic increase in integral dose difference for all points when the adjusted beam model is used

(**Figure 4.1A - B**). While this change in configuration caused slight variations in S_1 and Sp_1 the AUC was practically unchanged (See **Table 4.1**).

The same effect on measured dose occurs for the film and ArcCheck methods. However, this shift in absolute dose influences the γ pass rates in a different way for the two beam models (see **Figure 4.2**).

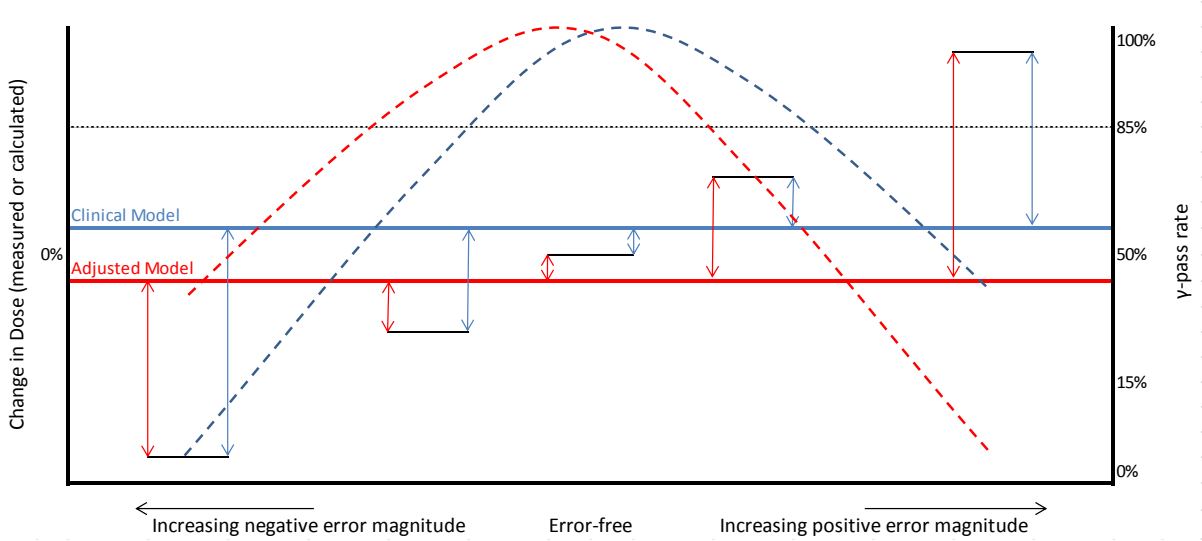

Figure 4.2: *Diagram indicating the effect of different beam models on γ pass rates. Blue and red lines represent the calculated TPS dose for the clinical beam model and the adjusted model, respectively. The black lines represent the measured dose for different error magnitudes. Arrows represent the dose difference (left y axis) for each combination of measurement and beam model. The resultant γ pass rates (right y axis) are displayed as dashed lines. The current 85% γ pass rate acceptance criterion is displayed as a black dashed line.*

The smaller DLG in the adjusted beam model will lead to a lower dose being calculated by the TPS. Therefore, when measured doses are low compared to calculated doses a dose difference map using the clinical beam model, will give better agreement between measured dose and dose calculated using the adjusted beam model (and therefore a higher γ pass rate). The inverse occurs when measured dose is high compared to calculated dose (and hence a low γ pass rate is obtained using the adjusted beam model). This is what is observed for both film and ArcCheck method QC results (see **Figure 4.1C - F**).

For the film results, the impact of the change in configuration is similar for the D_{98} and the D_{1cc} and the resulting change in the AUC of the ROC curve is limited from 0.77 to 0.80. However for the ArcCheck, the additional systematic dose offset has only a limited impact on D_{98}, but considerably influences the D_{1cc}. Therefore the change in the AUC from 0.74 to 0.60 when moving from the clinical model to the adjusted model is relatively large change compared to the change for the film QC method.

Therefore, sensitivity and specificity of patient-specific QC is not just determined by the properties of the measurement system, but the effects of beam modelling in the TPS need to be carefully considered. Furthermore the effects of the beam model should not be measured only for treatment plans that are considered acceptable, but also for unacceptable treatment plans. This is particularly important for QC methods where γ analysis is utilised as this may mask dose differences by compressing an entire 2D (or 3D) analysis into a single number representing the percentage of passing pixels (or voxels), which does not indicate why pixels are failing or how close passing pixels are to failing.

4.2.3. Influence of OAR Specific Intentional Errors on the Efficiency of Patient-Specific QC

For trPD and film measurements at the PTV locations, exclusion of intentional errors that specifically impact the dose delivery around an OAR increase the observed S_1 and Sp_1 because these intentional errors introduce changes in dose around the PTV that are smaller than the applied trPD acceptance criterion of 2% and γ-criterion of {2%;2mm}. These errors are therefore per definition not detectable at the PTV locations.

Remarkably, exclusion of the OAR specific intentional errors does not improve the results for the ArcCheck. This is because the ArcCheck effectively verifies the fluence of the dose delivery and

creates a 2D dose map of observed deviations at the surface of the ArcCheck phantom and there is no 1-to-1 relation between the dose to a PTV or OAR and the dose distribution on the ArcCheck dose map. No distinction is made between the dose delivered to the PTV or OARs in this verification method and OAR specific errors are therefore treated on an equal footing as any other error. Therefore, exclusion of OAR specific MLC errors probably only excludes errors which the ArcCheck can successfully detect.

4.2.4. Comparison of the Efficiency of WBCC Patient-Specific QC with Other Studies

The reported efficiency of patient-specific QC in literature varies considerably depending on the applied QC method. Unfortunately, it is not possible to make a fair comparison between the results of the different studies because there is a large variation in approach (intentional errors or otherwise, measurement locations), as well as a wide range of applied acceptance criteria. Our results show that these factors have a large impact on the observed efficiency of a QC method. For instance, for both point dose measurements using a single detector and for film dosimetry, the results are very different between measurements made at PTV or OAR locations, while the approach of the ArcCheck system does not allow differentiating between specific measurement locations. The studies by Kim et al. [44], Fredh et al. [49], Bojechko et al. [45], Yan et al. [83], and McKenzie at al. [85] included measurements at PTV locations only, or used global dose normalisation for deviations in the γ-analysis. This approach practically excludes OAR errors even if a low threshold such as 10% is used as a cut-off.

Furthermore, the studies investigating the sensitivity and specificity of the QC methods using intentional errors generally apply a) considerably larger errors than used in our study, as well as b) wider acceptance criteria. Looking at these two factors individually:

a) Increased error magnitude is expected to increase the values obtained for S_1, Sp_1 and AUC. This was also observed by Fredh et al. [49], Bojechko et al. [45] and Yan et al. [83]. A

comparison of our result for the smallest and largest error magnitudes showed significantly higher S_1 and Sp_1 for the larger error magnitudes. The difference in AUC between error magnitudes is larger than the 10% uncertainty margin for both MU errors and MLC open/closed shift errors (see **Figure 4.3**).

b) Tighter acceptance criteria are expected to result in higher S_1 values, but might also decrease Sp_1 considering that this increases the probability of FP results.

Although the overall effect of larger error magnitudes in combination with wider acceptance criteria is not readily made, it is logical to assume that larger error magnitudes would be detected by the QC methods at the WBCC given the increased efficiency for the largest error magnitudes investigated (see **Figure 4.3**). However, the efficiency of the methods while utilising wider acceptance criteria requires further investigation for direct comparison with other studies.

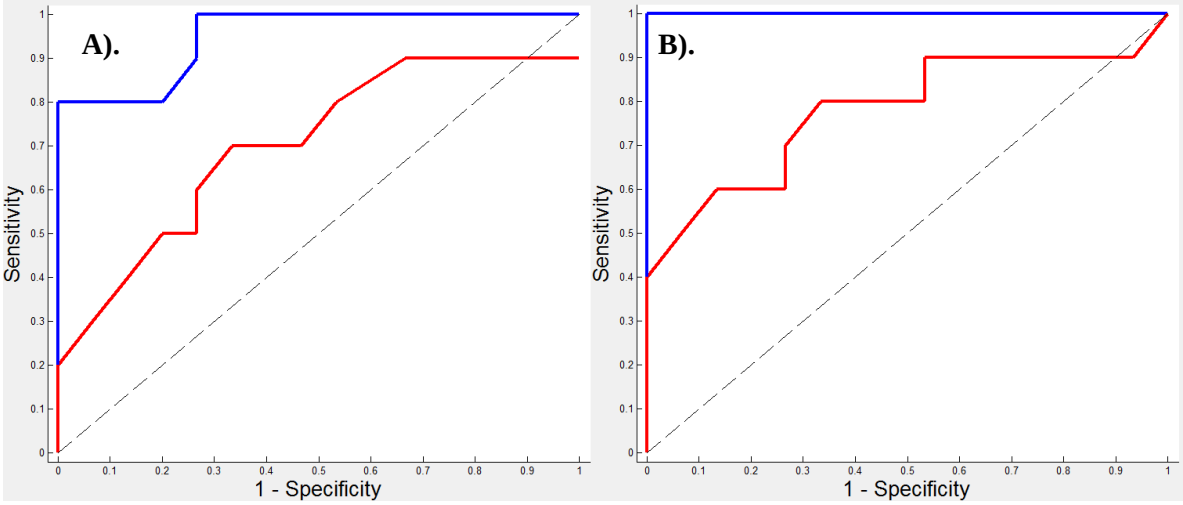

Figure 4.3 A-B: *ROC curves comparing error magnitudes for trPD measurements at PTV locations only.* ***A).*** *displays ROC curves for MU errors only. The AUC for ±3% MU errors (blue line) was 0.95 while the AUC for ±1.5% errors was 0.70 (red line).* ***B).*** *displays ROC curves for MLC open/closed shifts only. The AUC for ±1 mm shifts was 1.0 (blue line) while the AUC for ±0.5 mm shifts was 0.76 (red line).*

Other studies which didn't use a ROC-type analysis also showed a low sensitivity of patient-specific QC to detect errors. Kruse [52] indicated that there was no difference in either {3%; 3mm} or {2%; 2mm} γ pass rates for dosimetrically acceptable and unacceptable IMRT plans using an EPID and an

ion chamber array for patient-specific QC. Heilemann et al. [61] introduced 0.5 and 1.0 mm systematic MLC open and closed shifts to H&N plans and found only a small reduction in {2%;2mm} and {3%;3mm} γ pass rates for the 0.5 mm shifts while using two different arrays for QC measurements, although 1.0mm shifts could be detected by the {2%;2mm} γ-criterion. This was very similar to the results of the current study with respect to systematic MLC shifts.

Although all these intentional error studies have a large variation in experimental approach and applied QC acceptance criteria, a general trend appears to be that patient-specific QC shows a low sensitivity for detecting errors. In particular, the work of McKenzie et al. [85], Kry et al. [47], Aristophanous et al. [86], and Fredh et al. [49] is similar to the results for configuration **A** in the current study in that they indicate patient-specific QC methods generally seem to have high specificity but low sensitivity. It appears to be a common trend that patient-specific QC methods are optimised for passing acceptable QC plans as opposed to detecting plans which contain errors.

4.3.Intrinsic Sensitivity of the QC Methods Characterised by S_2 and S_3

The metrics S_2 and S_3 both quantify the change in 'output' relative to the change in 'input' of the QC method, and in that way characterise the intrinsic sensitivity of the QC methods. For S_2, the change in input is defined as the change in TPS calculated dose whereas for S_3, the change in input is defined as the change in error magnitude. These metrics are discussed for the trPD and film QC methods in this section, while all ArcCheck results are discussed in section 4.4.

trPD Results

S_2

The change in output is plotted against change in input for the trPD QC method (**Figure 4.4**). For the point dose method at PTV measurement locations (blue line, **Figure 4.4**), a near-perfect correlation is

observed (R^2 = 1.00), for change in output over change in input. This means that the point dose measurement can be directly correlated with what the calculated TPS difference is, and indicates that the point dose measurement method is highly sensitive. For measurements at OAR locations (red line, **Figure 4.4**), a slightly weaker correlation than that seen for the PTV locations is observed (R^2 = 0.95). This is likely due to the inaccuracy in positioning of the detector in the high dose gradients at the given OAR measurement locations.

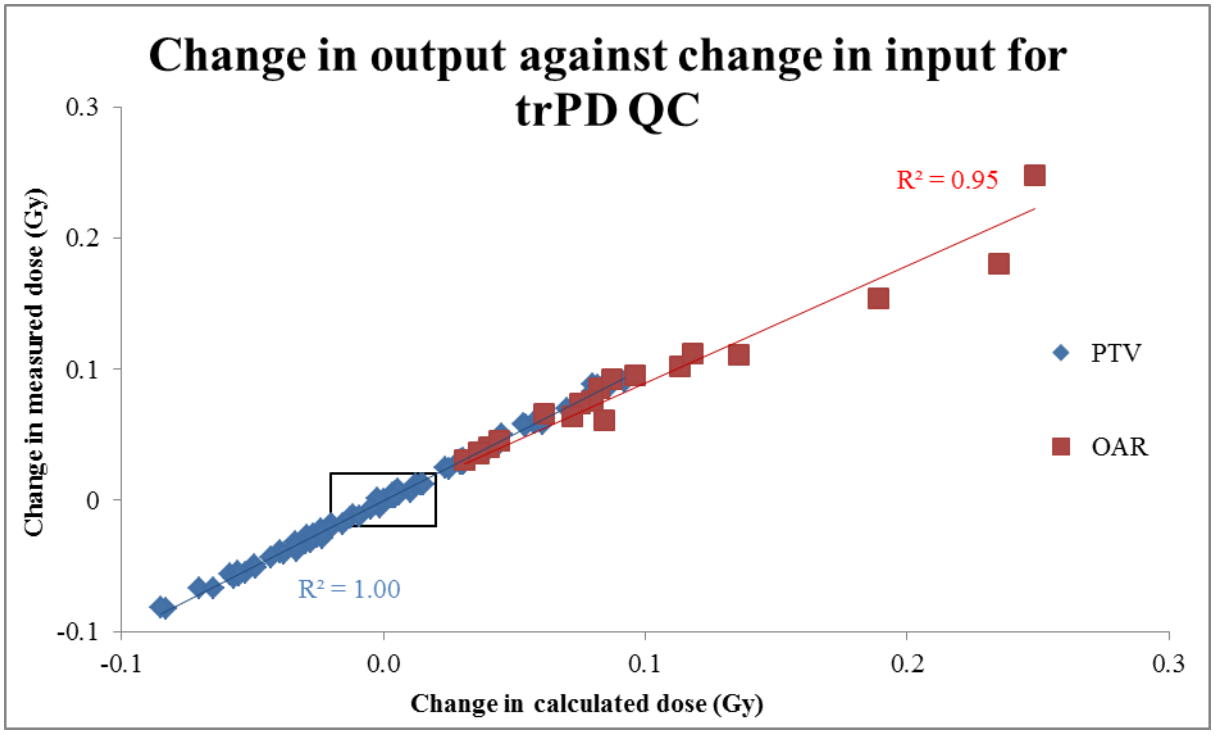

Figure 4.4: *Plot of change in measured % fraction dose deviation (system output) against change in TPS calculated % fraction dose deviation (system input) for all errors using the point dose method for the clinical beam model. Blue markers (and the blue trend line) represent PTV measurement points while red markers (and trend line) represent OAR measurement locations. The black box represents the region where the S_2 calculation becomes inaccurate.*

For the majority of errors the values of S_2 were 100 % ± 36 % (median ± 95th percentile of all data). This indicated that there is a one-on-one relationship between calculated and measured dose deviations for the trPD method. There were a number of error types which resulted in S_2 values outside the 95th percentile of all data (see also **Table 3.9**):

1) 1 mm MLC translation shift error

The change in TPS dose between the error and error-free plans was very small (mean change of 0.2 % ± 0.7 % [1 SD]). This is smaller than the measurement reproducibility. Therefore a small change in measured dose of the order of 0.1 % can lead to a large change in the calculated sensitivity. This explains why the sensitivity for the MLC translation shift error is over 100% and is well demonstrated by the results for patient one, which had a measured dose difference of 0.3 % and TPS dose difference of 0.07 % resulting in an S_2 of 416 %.

2) All OAR specific MLC shift errors

For all OAR specific errors the change in TPS dose at the PTV location is generally very small but more importantly: a 1-on-1 relationship would not be expected if the measurement point is located at a different position then the point where the error is introduced. For all OAR specific MLC errors, S_2 was closer to 100% when measuring at the OAR locations compared to PTV locations (see **Table 3.10**).

3) Combined 1 mm MLC closed shift and 3% MU increase errors

The median value for S_2 was 83.3 % with a wide range of values across the five patients. The two introduced errors are going to cancel one another out to a certain extent, leaving a small net dose difference compared to the clinical plan. If this net dose difference is close to zero, the resultant value for S_2 can be either positive or negative if small set up errors occur for the trPD measurements and will be observed as noise (similar to what was observed for the 1 mm MLC translation error). It should be noted that this plan was included as a worst case scenario of errors that will cancel one another out.

S_3

For MU errors: the linear regression for plot of trPD results against change in percentage of MU error introduced resulted in a strong correlation (**Figure 3.5**, $R^2 = 1.00$) and trend line with a slope of 1.0. As the MU change is equivalent to an overall dose change, S_3 is effectively the same metric as S_2 for MU errors.

For MLC shift errors: the linear regression for the plot of trPD QC results against change in magnitude of systematic MLC shift also showed a strong correlation (**Figure 3.6**, R^2 = 0.95). However there was more inter-patient variation compared to the plot for MU errors. This was due to the variation in the impact of MLC movement between each patient. Similarly as for S_2, measurements at PTV location for OAR specific errors yield different S_3 values overall compared to the systematic MLC errors, and a larger variation in S_3 values, and indicated that high S_3 can only be achieved for these OAR specific errors by measuring at the given OAR location.

Conducting a linear interpolation using the trend lines in **Figure 3.5** to the point where the error magnitude would become detected using the current QC acceptance criterion i.e. a 2% change in output, indicates at what magnitude an error is detected. For MU errors, a 2% change in MU is the threshold for detection which is the same as the tolerance on routine linac output checks [26]. For systematic MLC open or closed shifts, a 0.5 mm shift is the threshold for detection, which is well below the manufacturer's tolerance for the maximum deviation in individual leaf positioning [61], and is similar as the value reported by Kerns et al. [59] for MLC positioning accuracy of 0.5 mm for VMAT deliveries as reported using linac log files.

Film Results

The S_2 analysis was not carried out for the film dosimetry method as the current film analysis software did not enable comparison of two measured films. Therefore, the software would need to be modified which would have been too time consuming within the limited time constraints of this study and other analysis methods were prioritised instead. Furthermore, as stated earlier the variation in film response may impact on this analysis. The investigation into this effect is still on going, but the variation was observed in this study. This was evident from the variation between in two calibration films (from the same film batch) conducted during this study. The mean global dose variation

between these two films in the dose region of around 2.1 Gy (similar to the PTV dose for all patients) was found to be 1.4% (range 0.9 – 2.0%). This is comparable to the calculated change in dose at the PTV reference point (for example, the change in dose for the 1.0 mm MLC closed shift relative to the error-free plan for patient 1 was 1.5%). Therefore, the intra film batch variation in the film response was likely to provide a similar order of magnitude of random uncertainty as the observed change in QC results.

The S_3 results for film dosimetry display a larger variation for each error mode compared to the trPD QC results. This is likely due to two reasons:

- Firstly, the change in γ pass rates is dependent on the difference between the measured dose relative to the TPS calculated dose at each pixel location for the original plan. If these differences are close to zero for an error-free plan, an introduced error that results in a change in dose of less than the applied tolerance criteria will not change the γ pass rate at all. If the original difference between measured and TPS calculated dose is high, a large change in γ pass rate (and hence S_3) may occur even for small introduced errors.
- Secondly, the observed variation in film response has likely contributed to the large range of observed S_3 values. This is further discussed in section 4.5.2.

4.4. ArcCheck Results

S_3

The ArcCheck gave unexpected S_3 results; we would have expected S_3 to be negative for all errors indicating a reduction in pass rate whenever and error was introduced for the WBCC configuration. However, although S_3 was typically large and negative when errors were introduced that increased the delivered dose, it was small and also sometimes positive for errors that reduced the delivered dose.

The situation was worse for the recommended configuration where S_3 was positive for all the errors that decreased the delivered dose.

Looking more closely at the effect of the ArcCheck configuration on S_3, for both configurations, the measured dose was higher than the predicted dose, with the measured dose being the highest for the recommended configuration (when the HUo and HC are applied). Consequently (see **Table 3.31**, and **Figure 4.5** and **Figure 4.6**):

- all errors that lead to a reduction in dose (MU 1.5% and 3% decreases, MLC 0.5 mm and 1 mm closed shifts and both output variation with gantry angle errors) show an increase in pass rate (as the combination of the introduced error reducing the dose with the systematically high measurement moves the measured dose closer to the prediction) and consequently a larger positive increase in S_3 for the recommended versus WBCC ArcCheck configuration.
- all errors that lead to an increase in measured dose (MU 1.5% and 3% increases, MLC 0.5 mm and 1 mm open shifts) show γ pass rates falling, with a larger fall and corresponding decrease in S_3 for the recommended versus WBCC configuration.

Contrary to initial expectations, the poor γ pass rates observed for the ArcCheck were not caused by using the WBCC configuration rather than the recommended configuration i.e. not applying an HU override and heterogeneity correction. In fact, it appears that by not applying these corrections, the full extent of the systematic dose offset observed for the ArcCheck was partially masked by the slight dose decrease caused by not implementing these corrections. Therefore, further investigation is required to determine what is causing the observed dose offset; this could be some form of systematic error in the ArcCheck QC set up, measurement method, with the ArcCheck device itself, or in the calculation of the predicted dose.

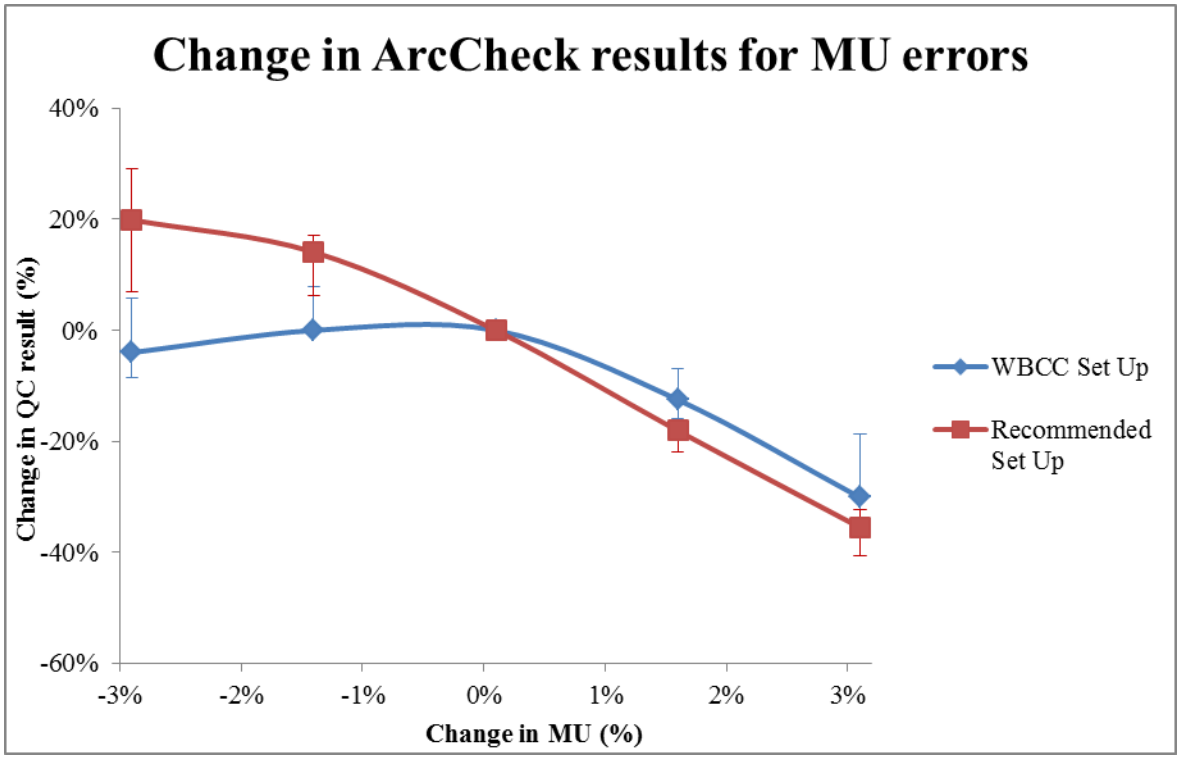

Figure 4.5: *Median value for change in ArcCheck QC results against systematic MU error magnitude (error bars display the range). The trend for the current WBCC set-up is given in blue, while the trend for the manufacturer's recommended set-up is given in red. A larger increase in pass rates for decreases in MU is observed using the recommended set-up relative to the WBCC set-up.*

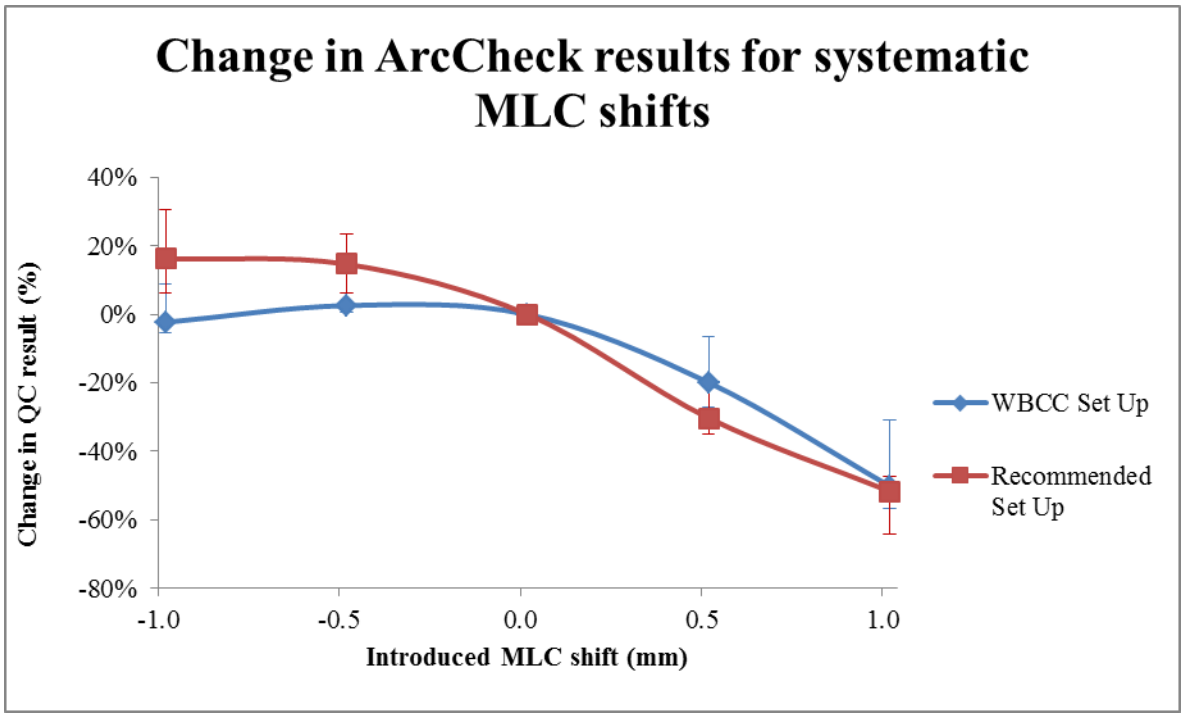

Figure 4.6: *Median value for change in ArcCheck QC results against systematic MLC error magnitude (error bars display the range). The trend for the current WBCC set-up is given in blue, while the trend for the manufacturer's recommended set-up is given in red. A larger increase in pass rates for MLC closed shifts is observed using the recommended set-up relative to the WBCC set-up.*

S_1 and Sp_1

This observed systematic measured dose offset as described above had a large impact on the efficiency of the ArcCheck QC method; it increased the rate of FPs at the expense of TNs for error-free and non-clinically relevant errors and also increased the rate of FNs at the expense of TPs for errors which caused a reduction in D_{98}. These two factors reduced S_1, Sp_1 and the AUC and resulted in none of the investigated ArcCheck configurations providing both acceptable S_1 and Sp_1 (see **Table 4.1**).

S_2

In terms of S_2 (see **Figure 4.7**) there was a strong correlation between change in output and change in input ($R^2 = 0.96$) for the ArcCheck method.

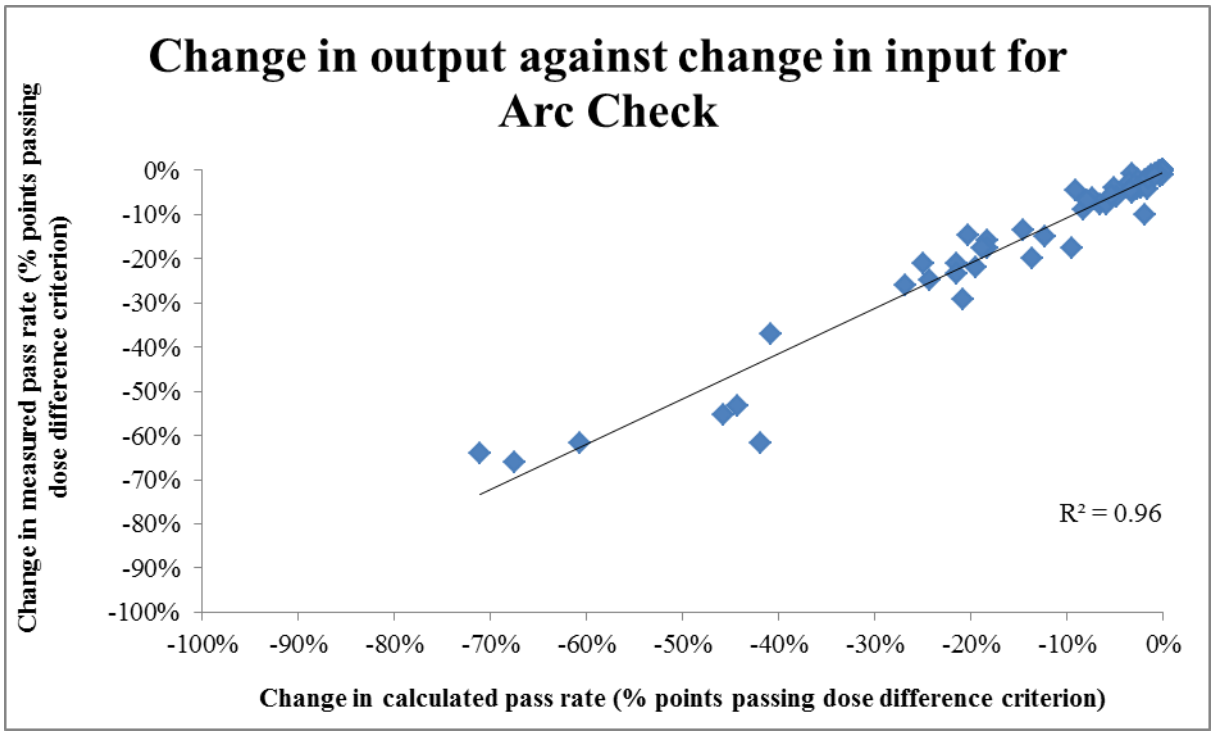

Figure 4.7: *Plot of change in measured percentage of points passing a {2%;2mm} γ-criterion (system output) against change in TPS calculated percentage of points passing a {2%;2mm} γ-criterion (system input) for all errors using the ArcCheck method for the clinical beam model.*

This result is conflicting with the poor S_1, Sp_1 and S_3 results observed for the ArcCheck. There are a number of factors that could account for this:

1) The S_2 analysis compared two measurements to obtain the change in output and two TPS dose maps for the change in input. If there is an error or dose offset that occurs during the measurement process (or alternatively occurs in the TPS), this will be applied to both measurements (and/or both calculations) and will not affect the resultant S_2 values.
2) These comparisons were conducted using a global dose difference criterion, meaning that any point between the verification plans must differ by more than 2% (or 3% depending on which criterion is used) of the maximum dose in the dose map. Therefore, only points where the dose is close to maximum are likely to fail a dose difference criterion.
3) The geometry of the ArcCheck may affect how the error appears in both the measured and calculated data. Most error modes are focussed on varying the dose to the PTV, which is located at or near the isocentre of the plan. However, the ArcCheck measures the radiation fluence in a helix with a diameter of 26.6 cm surrounding the isocentre, but does not measure near the centre of the volume. This may have the effect of 'blurring' out the error over the course of each arc and reducing its effect on any individual point.

Therefore the dose offset observed in the ArcCheck S_3 results did not affect the S_2 analysis. However, the ArcCheck results for S_3 indicate there is a dose offset present, and this dose offset can account for the low S_1, Sp_1 and AUC values observed for the ArcCheck. This dose offset issue is currently under current investigation, and the preliminary results from these investigations indicate that this issue may be partially due to a limitation in beam modelling of the TPS as well as a larger contribution of scattered dose from the accelerator head to the detectors compared to other QC methods.

4.5. QC Method Limitations

There are a number of limitations that are specific to each individual QC method, and these are discussed in the following subsections. However, there are also some limitations that apply to all the patient-specific QC methods tested in this study. Some of these can be minimised, but are always going to be present to some degree, such as phantom set-up variations (although these could be further

reduced by using image guided phantom set-up), or positional uncertainty limitations of the linac (e.g. gantry angle, collimator angle etc.). However, one limitation that applies to all QC methods which is more difficult to minimise is that all these patient-specific QC methods compare the measured dose to the TPS, and in doing so, it is difficult to determine in which part of the system an error has occurred. In a worst case scenario, an error in the TPS compensates for a delivery error resulting in an acceptable QC measurement. While this can be reduced by accurate commissioning and QC for the TPS and linac separately, an independent method of measurement should still be carried out in the form of routine external audits. Generally, an external audit has looser acceptance criteria than what is used clinically [47], so if a patient-specific QC method fails an external audit it should be extensively reviewed to determine why the QC method has failed.

4.5.1. Point Dose Limitations

The point dose measurement technique generally provided the most consistent results and was quite robust to changes in beam modelling. The main limitations of point dose measurements are that they are:

1) Only a dose measurement at a single point and a single point is not representative of the dose to an entire volume
2) Not able to detect errors that have little effect on the dose at the measurement point but which may have a clinically relevant effect elsewhere.

This has been demonstrated from all results for OAR specific MLC shift errors. When measuring at the appropriate OAR location, the point dose measurements have a very high S_1 and Sp_1 for these errors.

It is possible to mitigate this limitation by measuring a given plan at multiple locations but this has a number of drawbacks:

1) It is much more time consuming (as each change in position requires re set-up of the phantom) and even if multiple locations are measured they might still not be the right locations to detect a localised error.
2) Measurement positions are ideally limited to regions where the dose gradient of the integral fraction dose is relatively constant in the vicinity of the measurement point to limit the impact of positioning accuracy.
3) Measurement at OAR locations will likely not meet requirement (2) above; all the selected OAR locations had much larger dose gradients (up to 7.6%.mm^{-1}) compared to the PTV locations. Large dose gradients may cause a sizable disagreement between measured and calculated dose if a small positioning error occurs. This could potentially be mitigated by utilising a cone-beam CT to ensure accurate set up of the detector at the correct measurement location.

4.5.2. Film Dosimetry Limitations

Film measurements overcome the two main limitations associated with point dose measurements; they provide dosimetric information for an entire plane of the irradiated volume and the measured dose plane can be accurately positioned against the TPS dose plane (the film is manually registered to the TPS dose plane during analysis). Therefore a combination of point dose measurements and film dosimetry form a strong foundation for a patient-specific QC program as long as the point dose POI is within the corresponding film plane. However, film dosimetry does have limitations of its own: it is a more time consuming task than other QC methods, as the film needs to be exposed, left to develop for at least 18 hours, be scanned using a reproducible scanning protocol, and then be analysed. This provides more opportunities for random errors to add up. A significant limitation of the use of film at the WBCC was uncovered during this research, namely a larger than expected estimate on the uncertainty of the measurement of absolute dose of 1.0% (1 SD) compared to that of point dose measurements (±0.5 % [1 SD], this is excluding the uncertainty in the absolute dose as measured with the local standard).

This limitation which was uncovered during this research (and during departmental application of patient-specific QC using film) was due to the large intra-batch variation in film response. There are a number of factors that could potentially contribute to this such as phantom set up errors, variation in linac output during a measurement session, film scanner variations and variation in film response. However this variation appeared to be specific to the film QC method considering this level of variation was not seen in either of the other QC methods, leaving film scanner variations and variation in film response as the main potential sources of the discrepancies seen in the film analysis. A departmental investigation into uncovering the cause of this variation was conducted, which looked into the variation in the lateral scan effect observed between different calibration films within the same film batch for both irradiated films included in this study, and films irradiated as part of the WBCC routine QC program. The details of this investigation are outside the scope of this study, but the investigation uncovered a correlation between the magnitude of the observed intra batch film variation and the time between film manufacture and irradiation, with the variability reducing as the time between manufacture and irradiation increases. These facts were not known when the experiments were carried out for this MSc thesis study. Many of the film batches used for this MSc-thesis were procured for the sole purpose of the study, and only 1 - 2 calibration films of the same film batch were made within a single measurement session. Some film batches used for this study displayed an exceptionally large variation in film response in terms of variation of the lateral scan effect. However due to the very limited number of calibration films per batch, it is not possible to make a reasonably accurate estimate of the impact of film response variation on the results of this study. The departmental investigation is still currently on-going, but changes have already been implemented for routine film QC at the WBCC (see section 4.8).

4.5.3. ArcCheck Limitations

The ArcCheck is a commercially available QA tool that could potentially enable inter-departmental comparison of QC results. However over the course of this study the ArcCheck was found to have the lowest sensitivity and specificity over all metrics used for analysis out of all three investigated QC

methods due to a systematic dose offset which has been discussed in section 4.4. The ArcCheck is not currently used clinically for patient-specific QC at the WBCC.

4.6. ROC Analysis Limitations

The ROC analysis method used during this study also has a couple of limitations:

1) The reference standard (in this study this was whether the clinical plans were truly acceptable treatment plans) will contain its own uncertainty. However it would be very difficult to quantify this and for practical reasons the five clinical plans have been assumed to be acceptable clinical treatment plans in this study. This could in principle have had an effect of the ROC curves that were generated.
2) The AUC for each ROC curve in this analysis has been calculated simply by connecting each point of the ROC curve with straight lines and summing the areas of the resultant rectangular and triangular areas (known as the trapezoidal method). However this technique has been shown to systematically underestimate the true AUC based on a smooth curve [88]. A more accurate analysis could be conducted by determining the AUC using the two-sample Mann-Whitney rank-sum test [79 - 80].
3) The uncertainty of the ROC analysis method needs to be carefully considered. One method to do this is by estimating the 95% confidence interval for a given AUC using a bootstrapping procedure similar to that used by McKenzie et al. [85]. Due to time constraints, this was not conducted during the current study and an uncertainty of ±10% was assumed.
4) Finally, the analysis to determine the best cut-off criterion for determining the optimal sensitivity and specificity of each QC method was conducted by using the Youden index. This finds the point on the ROC curve which is furthest from the diagonal 0.5 AUC line that corresponds to a non-informative test. While this method is very simple to conduct, it implies that the impact of a false negative is equivalent to that of a false positive. Whereas in the context of patient-specific QC, a false positive may lead to an increase in staff workload to investigate the cause of the positive reading and conduct repeat measurements. On the other

hand, a false negative could lead to an incorrect patient treatment for one fraction, multiple fractions or an entire treatment course. Therefore, the outcome from a FN can potentially be much worse than for a FP. Thus, patient-specific QC should be weighted towards minimising FNs (and hence maximising sensitivity) at the expense of marginally increasing the number of FPs (or decreasing specificity). The extent to which this should be done is determined by optimising the cost function for FNs compared to FPs which should also take into account the expected rate of occurrence for errors in treatment plans. This has already been conducted for one previous study [85] and would be logical future work for this current study.

4.7. Resolving Error Modes

An important area of interest in this study was in defining specificity as the ability of a QC method to resolve different error modes as opposed to the true negative rate of the QC system. This was only attempted for one of the three QC methods (trPD measurements), although a method for resolving error modes using film dosimetry and the ArcCheck was also proposed (see section 4.9). The trPD method showed that it is feasible to resolve MU errors and systematic MLC shift error modes using a method that requires no extra experimental data acquisition and only a small amount of extra data processing. However, further work needs to be conducted before this methodology can be implemented into routine QC practise.

The advantage of being able to resolve error modes is not simply in being able to determine what type of error has occurred (as a follow up investigation into the failure of any patient-specific QC method should determine the cause of the failure), but because it provides an immediate insight into what might be the reason for the observed failure. From the initial QC result, it can be determined if it would be more beneficial to investigate the linac output, or MLC positioning, or if a repeat measurement should be conducted. It is proposed that by providing easily accessible information on why the failure has occurred that the next course of action will be to resolve that failure, as opposed to simply repeating QC until a result passes.

4.8. Recommendations

There are a number of recommendations for each of the individual QC methods investigated in this study.

- For the trPD QC method, the reduced specificity and sensitivity at OAR measurement locations was likely due to set-up errors when positioning the phantom. Therefore it is recommended that image-guidance using a CBCT should be used to more accurately position the phantom prior to trPD measurements.
- For the film dosimetry method, a large intra batch variation in film response was observed. No current solution has been identified to correct for this. However a workaround has been implemented to reduce the uncertainty in film QC results. This involves irradiating multiple calibration films in order to identify films where the response is different from the majority of calibration films. Measurement films are also compared to the corresponding point dose location in the film plane in order to determine if a film measurement is an outlier or not, (film measurements are rejected if the difference between film and point dose is more than 2%). As a final resolution it may be necessary to re-measure the plan if uncertainty remains in the film measurement. These recommendations have already been incorporated into the film dosimetry program at the WBCC.
- For the ArcCheck, the source of the systematic dose offset needs to be investigated and resolved before the ArcCheck is returned to clinical use.

More generally for patient-specific QC, it is not sufficient for a QC method to have a high AUC value for it to be considered appropriate for routine use, but the QC acceptance values should be appropriately set to optimise S_1 and Sp_1 taking into account the uncertainty associated with the ROC analysis method. As mentioned in sections 4.2 and **Error! Reference source not found.**, the results of this study and previous studies in the literature seem to indicate that although QC methods may have a reasonable AUC, the current passing criteria values are set such that the test will have a high Sp_1 but a low S_1. This is converse to the ALARA principle of reducing the risk of errors in patient treatments to the extent that is reasonably achievable. Furthermore, high specificity and low

sensitivity may result in a false perception of the outcome of patient-specific QC. For instance, if treatment plans always pass QC, and one measurement then fails QC, it may be perceived as an outlier and be re-measured until a measurement indicates the plan passes. Whereas if the QC method truly had low sensitivity, the treatment plan could potentially have a large error in it to fail the QC (assuming the QC failure was not due to a QC set-up error or other error in the QC measurement process).

The current international recommendations for patient-specific QC acceptance criteria are given in AAPM TG 119 report. They are ±5 % for point dose measurements and a γ pass rate of 88 – 90 % for a {3%;3mm} γ-criterion [89]. These recommendations were obtained using a statistical analysis from multiple institutions. However, they were based only on an achievable pass rates for QC methods using the analysis of correct plans and did not consider plans containing errors. Furthermore, these criteria are the same (or looser, meaning they are more likely to pass plans that contain errors) than the criteria used in this study, and other studies [47, 85], which demonstrated that patient-specific QC is insensitive to detecting errors. Studies by Nelms et al. [90] and Yan et al. [83] have already indicated that a {3%;3mm} γ-criteria provides insufficient sensitivity and tighter tolerances should be implemented. Our study further supports the recommendation to move away from the {3%;3mm} γ-criteria and towards a {2%;2mm} γ-criteria.

More generally, it is recommended that QC acceptance criteria should not only be based on analysis of acceptable treatment plans but that they should be based on the outcomes of intentional error studies. This will increase the workload of physicists conducting patient-specific QC (as more investigations into QC failures will be required), but will reduce the likelihood of unacceptable plans being used to treat patients. Furthermore, it is recommended that when a patient-specific QC method is being commissioned for clinical use, the ability of the method to detect errors should be quantified as part of the commissioning process.

4.9. Future Work

Patient numbers included in this study

Due to time constraints, 'only' 5 patients were included in this study to enable the investigation of 6 different error modes each with various error magnitudes, for 3 different QC methods. Although this yielded a wealth of information and a large amount of data, this limitation in patient numbers prohibited the application of statistical tests in a meaningful way. This also limited the ability to assess the uncertainty of the ROC analysis conducted in this study. Consequently, it was not possible to prove whether configuration changes in the TPS beam model and/or the QC acceptance criteria resulted in statistically significant improvements. A bootstrapping method for determining the uncertainty of the AUC has been conducted in a study by McKenzie et al. [85], and this may be a useful method to investigate the uncertainty of the AUCs for the results of the current study. Nevertheless, the next step in this investigation will be to expand the patient cohort to at least 8 patients to assess the statistical significance of the observed changes.

Resolving error modes using film dosimetry

An important aspect of this study was investigating whether different error modes could be resolved using patient-specific QC. This was investigated using the trPD QC method, but due to time constraints it was not investigated for either the film or ArcCheck QC methods. The following is a potential method that could be used to resolve error modes for the film QC method. This is similar to methodology used to investigate the ability of the trPD method to resolve error modes (see section 2.8.2). But instead of using the DTFE to determine specific regions as for the trPD method, dose levels and dose gradient constraints could potentially be used to define three separate regions. The rationale behind these criteria are that a MLC shift is expected to mainly effect dose in the region near the field edge, which is the region of a dose map where the highest dose gradients are likely to be, while an output error is going to affect the open field, which is the region of the dose map where the dose is high, but the dose gradient is smaller. Once all the pixels of a dose map have defined to belong

in each of the three regions, gamma analysis can be performed per region to resolve error modes using similar definitions and criteria as defined for the trPD QC method. For example the following criteria could potentially be used to define the three separate regions:

$$\text{Region I:}\quad D < D_{lower} \text{ and } \Delta D/mm < (\Delta D/mm)_{crit}$$

$$\text{Region II:}\quad D_{lower} \leq D \leq D_{upper} \text{ and } \Delta D/mm > (\Delta D/mm)_{crit}$$

$$\text{Region III:}\quad D > D_{upper} \text{ and } \Delta D/mm < (\Delta D/mm)_{crit}$$

Where D and ΔD/mm are the dose and dose gradient over a particular pixel respectively. D_{lower} and D_{upper} are user defined thresholds for a low dose level and for a high dose level respectively, and $(\Delta D/mm)_{crit}$ is the user defined dose gradient value above which the pixel is deemed to be in a high dose gradient region.

This method is included as a proposed method to resolve error modes using the film QC method. Obviously, prior to initiating a feasibility study into this method, the reproducibility of the film response would need to be improved.

5. Conclusions

Patient-specific QC is a vital component of a radiotherapy department's quality management program, particularly if complex treatment techniques are utilised. However, it is not standard clinical practise to quantify the characteristics including the sensitivity (ability to detect errors) and specificity (ability to ignore irrelevant errors and pass acceptable plans) of the patient-specific QC method in detail.

This study involved the introduction of various intentional error modes and defined a metric to quantify the clinical relevance of an introduced error. All three of the WBCC patient-specific QC methods as clinically applied at the start of this study were found to have a relatively low sensitivity but high specificity. While the sensitivity of the trPD and film methods could potentially be improved at the expense of a small reduction in specificity, the observed improvement was within the estimated margin of uncertainty. Adjusting the TPS beam model had little impact on the overall efficiency (in terms of the AUC of the ROC curve) of trPD and film measurements. By investigating the intrinsic sensitivity of the ArcCheck, a systematic offset between the measured and TPS calculated dose was discovered which caused the ArcCheck's low efficiency. Several configuration changes were investigated to improve the ArcCheck efficiency but none of these changes provided an acceptable sensitivity and specificity of the ArcCheck system.

We have made recommendations on how to potentially improve the efficiency of each of the three QC methods investigated in this study:

- For the trPD method, the efficiency can potentially be improved by utilising image-guided phantom set-up to improve the positioning accuracy of the detector, specifically for locations with a high dose gradient of the 3D fraction dose distribution.

- For the film method, the observed intra-batch variation of the film response is a major limitation of the QC method efficiency, and improvements in film response reproducibility are urgently needed.
- For the ArcCheck, additional investigation is required to determine the cause of the systematic dose offset before the ArcCheck is returned to clinical use.

The results of this study are consistent with the findings of previous papers in literature which indicate that most patient-specific QC methods utilising commonly accepted QC acceptance criteria are not optimised for a high sensitivity but are optimised to have a high specificity. Furthermore, the number of studies quantifying both these characteristics is considerably smaller than the number of studies focussing on the ability of a QC method to pass acceptable treatment plans. Therefore, one can argue that *effectively* patient safety does not get the same priority as the manageability of the workload, even though the primary goal of patient-specific QC is to catch potential errors and improve patient safety. We therefore recommend that both sensitivity and specificity of QC methods are standardly characterised before clinical use and that QC acceptance criteria are optimised based on an ROC-type analysis. It should be noted however that this requires a considerable investigation in terms of both time and resources which may not be possible for individual departments, in particular to include a larger cohort of patients. A possible solution for this problem could be that departments with similar equipment share the workload of such a study.

This study has shown that it may be feasible to resolve different error modes using a trPD verification measurement of a single patient plan. This would enhance the efficiency of patient-specific QA by reducing the time required to investigate failed QC results.

6. Appendices

Appendix 2.A. Beam Model Optimisation Methodology

Since this project involved introducing intentional errors to the TPS beam model, it was desirable to ensure the initial beam model was optimised for the error-free treatment plans. The physical minimum DLG and focal spot of the treatment machine is different from the corresponding values set in the treatment planning system, and the value set in the linac control software in the case of the DLG (see sections 2.4.4 and 2.4.5). Therefore, it was necessary to determine which values of these parameters provide the best agreement between measured data and TPS generated data, as this may not necessarily be the measured value or the manufacturer recommended value [62, 91]. It was also important to consider that the optimal setting for these parameters may not be the same as for the beam model that was used to calculate the clinical plans. Of particular interest in the context of this research was ensuring that the ETSS and DLG were correctly set in the TPS. Two methods were used to determine the optimal beam model parameters. These included:

- Time resolved 'sweeping gap' measurements with a pinpoint ionisation chamber
- Using EBT3 gafchromic film to measure MLC defined fields and comparing these to TPS calculated data.

For each of these methodologies, measured data was compared to data calculated using several different beam models in which the DLG and TSS parameters were varied. **Table 2.A.1** shows the different values for the DLG and ETSS that were tested. Only two values of DLG were tested as these correspond to the current DLG setting in Eclipse for calculating treatment plans, and the DLG measured as part of routine QC of the linac to be used for this research. The ETSS was only varied in the x direction (direction of MLC leaf travel); time resolved sweeping gap measurements showed that the current ETSS Y setting of 0.0 mm matched the experimental data accurately (see appendix 3.A).

Table 2.A.1: *Parameters for the four different beam models tested to determine the optimal beam model*

Beam Model No.	DLG (mm)	ETSS X (mm)	ETSS Y (mm)	Notes
1	2.0	0.0	0.0	Current clinical beam model
2	1.2	0.0	0.0	
3	1.2	1.0	0.0	
4	1.2	1.5	0.0	Adjusted beam model

Time Resolved Sweeping Gap Measurements

A common method to determine the DLG of a linac photon beam was outlined by LoSasso et al. [92] and Mei et al. [93]. This method involved using an ionisation chamber to measure sliding window (SW) IMRT plans of various field sizes. The ratio of the ionisation chamber reading for the SW plan compared to an open field (also taking account of MLC transmission) was plotted against SW field size. Then by fitting a linear trend line, the DLG can be obtained by finding the x intercept of the trend line (when the ionisation chamber reading is zero). This test was used at the WBCC to determine the DLG of each linac beam during commissioning (and is used for routine MLC QA). By implementing this test using a time-resolved measurement, it was also possible to use this test to analyse the beam penumbra and investigate the effect of varying the ETSS parameters. A test IMRT plan was developed in Eclipse where a sweeping MLC window of 2.0 cm width moves across the isocentre in a 10 cm by 10 cm jaw defined field. Measurement and analysis was conducted using the same methodology as outlined in section 2.6.2. The results of this analysis are discussed in appendix 3.A.

EBT3 Gafchromic Film Measurements

Gafchromic film measurements were also made to verify the result of the time-resolved sweeping gap measurements. A test plan was created in Eclipse which consisted of a static MLC defined pattern of four squares as used by Louwe et al. [51] (see **Figure 2.A.1** for the shape of the pattern on an irradiated film). This MLC pattern was chosen as its symmetry allows a reduction in the uncertainty of the registration between the film and the TPS dose plane, which in turn reduces the impact of registration errors on the analysis. Furthermore, the different sizes of the squares allow measurements

of several MLC penumbrae at various off axis distances on the same film. This MLC pattern was used to irradiate a sheet of EBT3 film at a depth of 5 cm (95 cm SSD) in a plastic water slab phantom with 355 MU. The film was subsequently scanned and analysed using the method outlined in section 2.6.3. The results of this analysis are discussed in appendix 3.A.

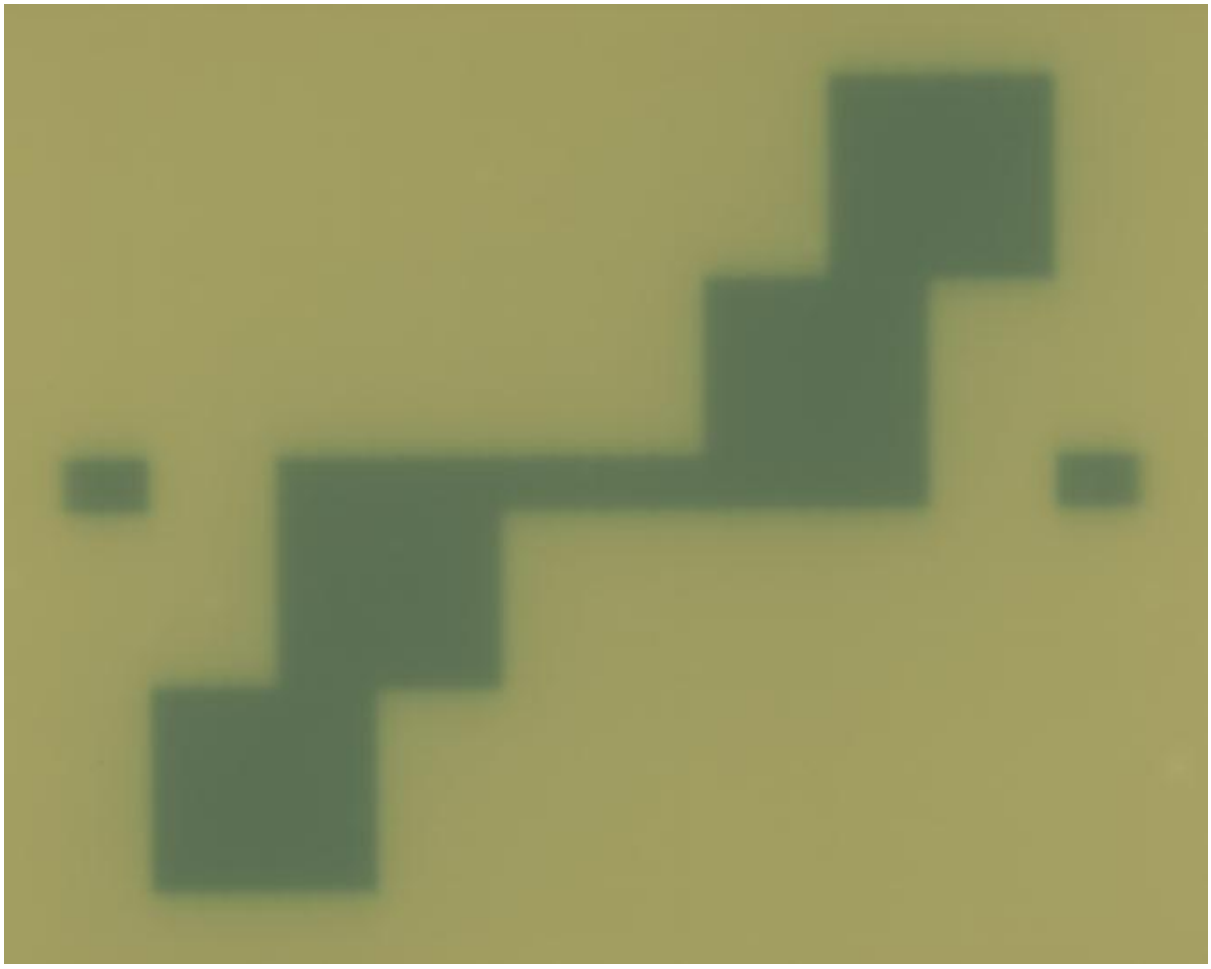

Figure 2.A.1: *Scanned film showing the 4 square MLC pattern used for EBT film analysis during beam model adjustment.*

Appendix 2.B. Software Development

Throughout the course of this study, a number of challenges arose which required the development of in-house software to provide a solution. All software development was carried out using Matlab (v2014a, The Mathworks, Natick, MA, USA), by generating a graphical user interface (GUI) which allowed the user to complete the particular task required. Overall, seven separate pieces of in house developed software were used during this study. Four of these were developed prior to this study for use in the WBCC routine patient-specific QC program.

Three of these programs are required for the trPD technique (see section 2.6.2). One program to allow for time-resolved signal acquisition using a PTW Tandem electrometer, one program to manage the signal acquisition files, linac log files, and DICOM RT files from the TPS and produce an input file to the analysis software, and one program to conduct the time-resolved analysis.

Another program was produced in order to carry out the film analysis method (see section 2.6.3). All of these software programs had been in routine clinical use for a number of years prior to this study and have been robustly tested.

The remaining three software programs were developed over the course of this study. Two of these programs were developed to allow the introduction of errors to treatment plans, one for linac output errors that varied with gantry angle (see section 2.5.2) and one for MLC positioning errors (see section 2.5.3). These software programs both directly import the original DICOM RT treatment plan, allow introduction of the relevant error mode, and then export a new DICOM RT treatment plan containing the error. These programs were verified by reimporting the plan containing the introduced error into the TPS and comparing the modified parameters with the original treatment plan to ensure the error was introduced as intended.

The final software program that was developed was used to conduct ROC analysis and AUC calculation (see section 2.9) for each QC method. This involved importing the QC results from a spreadsheet, calculating the sensitivity and specificity values (using S_1 and Sp_1 definitions) for a varying QC passing criterion, and plotting sensitivity against specificity. The accuracy of the calculation of S_1 and Sp_1, and the AUC calculations were checked against manual calculation for a limited number of points and ROC curves.

Appendix 3.A. Beam Model Adjustment Results

This section outlines the results of the beam model adjustment measurements that were conducted using the methods outlined in appendix 2.A. It also covers the analysis and conclusions about the parameters to be used for the adjusted beam model.

Time Resolved Sweeping Window Measurements

Time resolved sweeping gap measurements were made using the method outlined in appendix 2.A. These were then compared to the TPS calculated dose for the verification plan using the current clinical beam model settings which consisted of DLG = 2.0 mm and ETSS of 0.0 mm in both x and y directions. **Figure 3.A.1** below shows the dose per control point for the experimentally measured data compared to the TPS calculated dose.

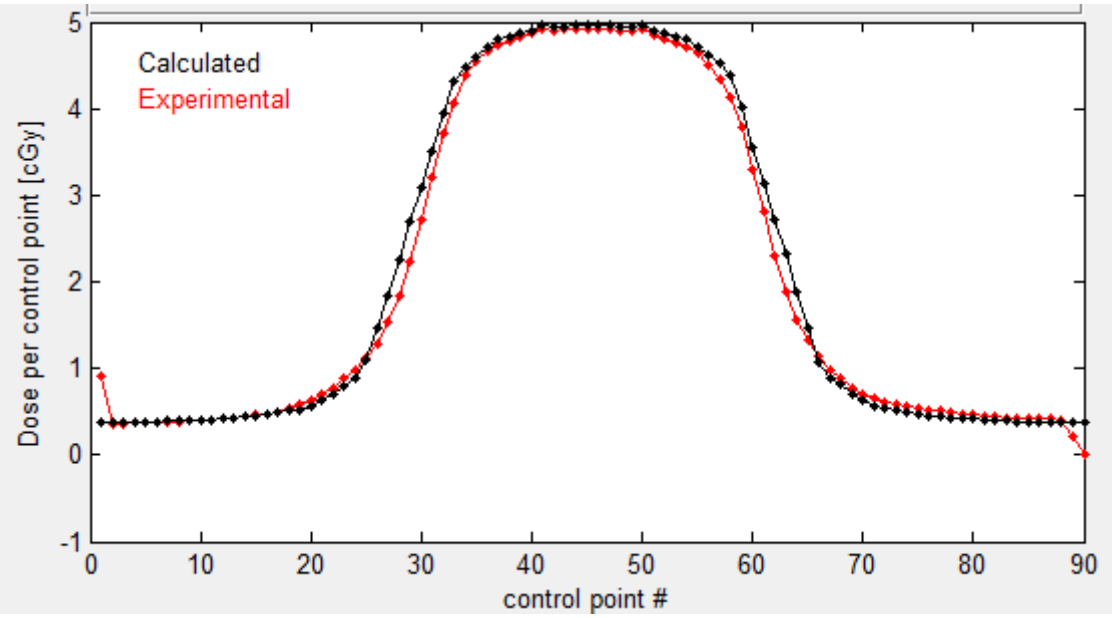

Figure 3.A.1: *Time resolved sweeping window analysis of clinical beam model (DLG 2.0mm, ETSS 0.0mm in x and y direction).*

From the above plot, it is very clear that calculated dose per control point is overestimating the dose in the vicinity of the rounded MLC leaf edge. Earlier in house investigations have indicated that the DLG setting effects the width of the calculated dose distribution (i.e. the distance between the 50% dose points) while the ETSS settings effect the rounding off of the penumbra at both the shoulder and tail of the penumbra [94]. This indicated that the DLG is currently set too high based on the

experimental data. The same measurement was then compared to the TPS calculated dose with the DLG set to 1.2 mm (ETSS still 0.0 mm in both x and y directions). This is the minimum DLG value that was measured on the linac during commissioning.

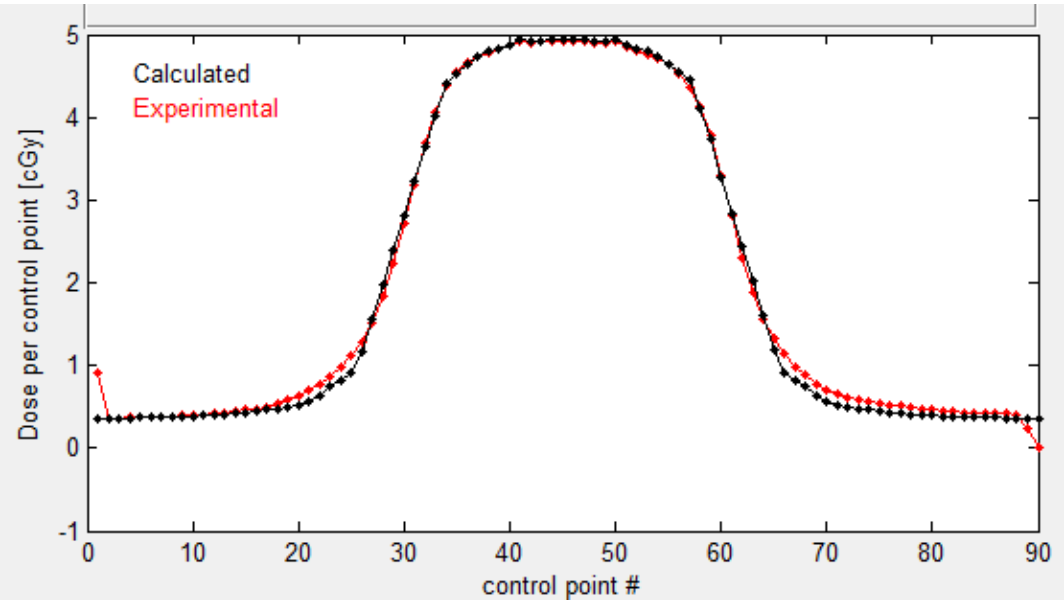

Figure 3.A.2: *Time resolved sweeping window analysis with the DLG set to 1.2 mm (ETSS 0.0 mm in both x and y directions).*

Calculating the dose distribution using the measured value of the DLG (1.2mm), it can be seen from **Figure 3.A.2** above that the experimental and calculated doses match better in the vicinity of the MLC leaf edge. There is still a mismatch in the penumbra where the experimental dose distribution has a rounder dose distribution that calculated. However, this is due to the ETSS settings as opposed to the DLG setting.

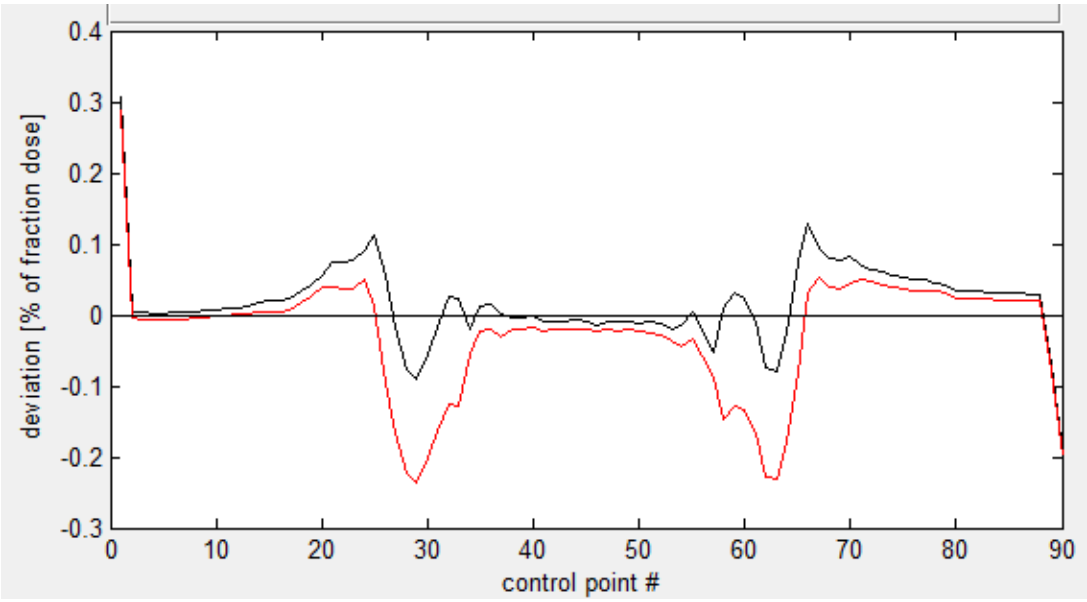

Figure 3.A.3: *Deviation between measured and calculated dose per control point for the sweeping gap measurements for the two different DLG settings. DLG = 2.0 mm is shown in red and DLG = 1.2 mm is shown in black.*

Figure 3.A.3 above shows the percentage deviation per control point between the measured and calculated dose for both DLG settings. It is obvious that the TPS calculation with a DLG setting of 2.0 mm causes a systematic underdose to be measured in the penumbral region, while the TPS calculation with the DLG set to 1.2 mm is centred on zero deviation with smaller positive and negative deviations at the tail and shoulder of the penumbra respectively.

A sweeping gap measurement was made in the y direction with the ETSS in the Y direction set to 0.0 mm (see **Figure 3.A.4**). This analysis showed that the ETSS Y setting in the TPS of 0.0 mm fitted the measured data very well. Therefore, this value was selected as the optimal ETSS Y value and was used for all calculations and no other values for the ETSS in the Y direction were investigated.

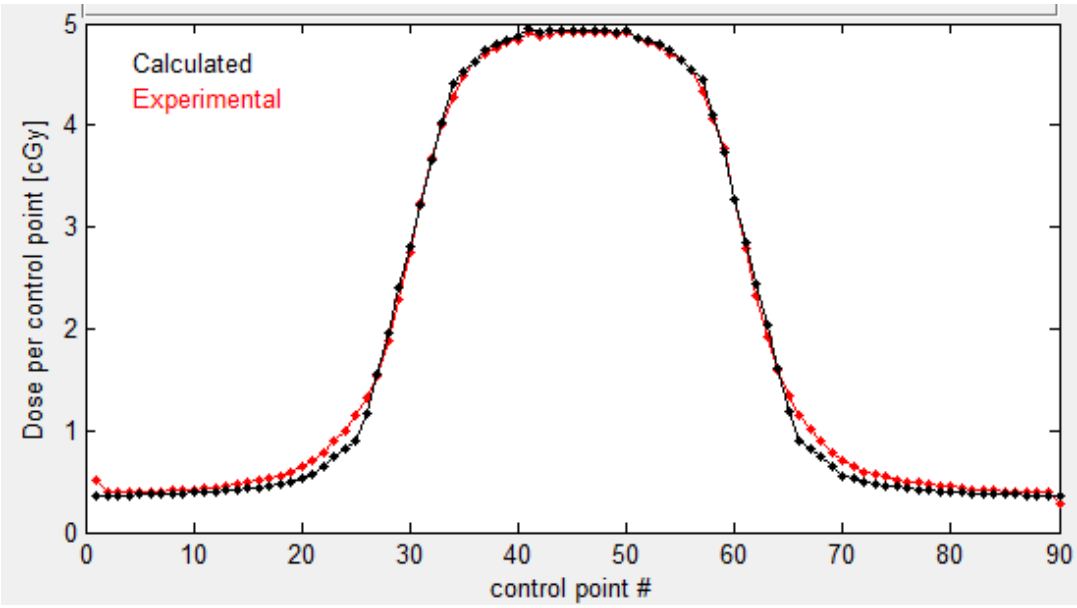

Figure 3.A.4: *Time resolved sweeping window analysis in the Y direction of with ETSS set to 0.0 mm in both directions (DLG = 1.2 mm).*

Subsequently, three different values of the ETSS in the x direction were used to calculate the dose for the sliding window plan, and were compared to the measured data.

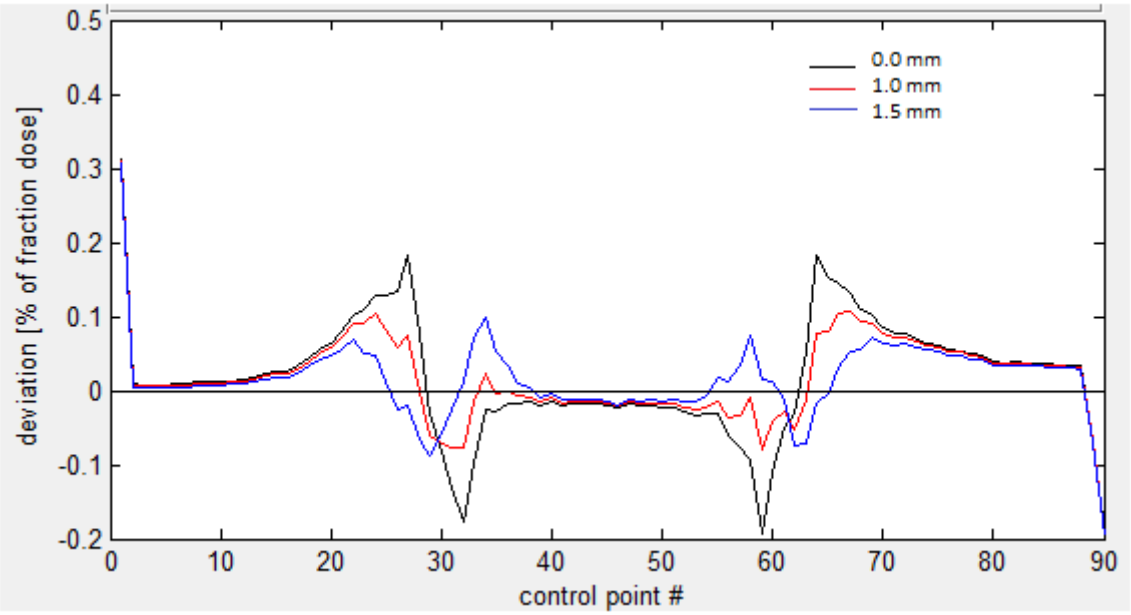

Figure 3.A.5: *Deviation between measured and calculated dose per control point for the sweeping gap measurements for the three different ETSS X settings. ETSS X = 0.0 mm is shown in black, ETSS X = 1.0 mm is shown in red and ETSS X = 1.5 mm is shown in blue.*

For the data displayed in **Figure 3.A.5**, it can be seen that all comparisons result in the familiar pattern of small positive deviations in the tail of the penumbra and negative deviations in the shoulder.

However, it is fairly obvious that the ETSS of 0.0 mm (black line) in the x direction gives the worst agreement with measured data. The difference between the deviations for the 1.0 mm and 1.5 mm ETSS X are much more similar. The 1.0 mm ETSS X results (red line) in a larger deviation in the tail of the penumbra but a smaller deviation in the shoulder compared to the 1.5 mm ETSS X (blue line). It appeared that the 1.5 mm ETSS provided a better fit, as the positive and negative deviations in the tail and shoulder respectively are more similar in magnitude, whereas the positive deviation in the tail is larger than the negative deviation in the shoulder for the 1.0 mm ETSS X. This means that for a VMAT plan the errors introduced by the 1.5 mm ETSS will more or less cancel each other out, whereas the 1.0 mm ETSS will result in a small net increase in dose deviation over the entire arc.

Therefore, the sweeping window measurements indicated that the optimum beam parameters were a DLG of 1.2 mm and an ETSS of 1.5 mm in the x direction.

Film Measurements

EBT3 film measurements of the static 4 square MLC plan were made using the method outlined in appendix 2.A. These were then compared to the TPS calculated dose for the verification plan using the current clinical beam model settings which consisted of DLG = 2.0 mm and ETSS of 0.0 mm in both x and y directions.

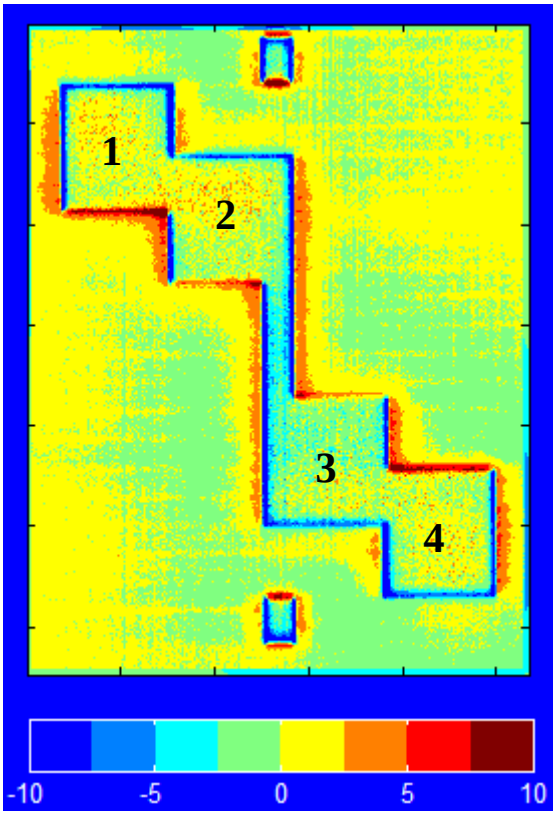

Figure 3.A.6: *Global dose difference display for the irradiated 4 square MLC film compared to the TPS calculated dose using the clinical beam model (DLG 2.0mm, ETSS 0.0mm in x and y direction). The numbering in each MLC square will be used to refer to each square the continued analysis below. The horizontal (x) direction is parallel to the MLC leaves, while the vertical (y) direction is perpendicular to the MLC leaves.*

Figure 3.A.6 above shows the global dose difference plot between the irradiated film and the TPS calculated dose. A γ analysis (absolute dose, {2%;2mm} criterion, 50% threshold) indicated a pixel passing rate of 94.2%. However, it is fairly obvious that there is poor agreement in the edges of each square. As the film analysis software did not allow sub millimetre shifts, the data had to be exported from the analysis software into an Excel spreadsheet, where the profiles from each of the 4 squares were analysed manually. This allowed sub-pixel global shift of all film data relative to the TPS data. The film data for each MLC square could then be translated individually to provide the best fit with TPS data. Finally, plots through each MLC square in both the horizontal (x) and vertical (y) directions were made for the dose difference between measured and TPS calculated dose.

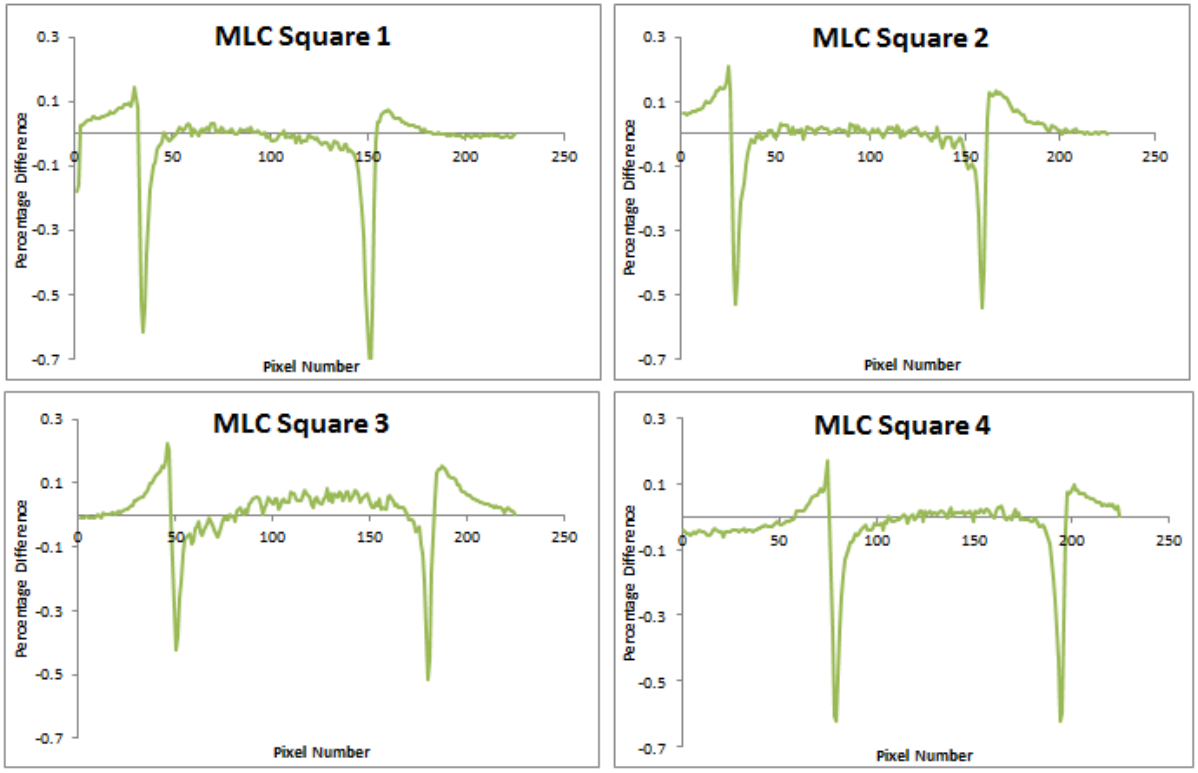

Figure 3.A.7: *Comparison profiles through each of the four MLC squares in the x-direction showing the difference between measured film dose and TPS calculated dose using the clinical beam model (DLG = 2.0 mm, ETSS = 0.0 mm in both x and y directions). The y axis of each plot shows percentage difference while the x axis is pixel number.*

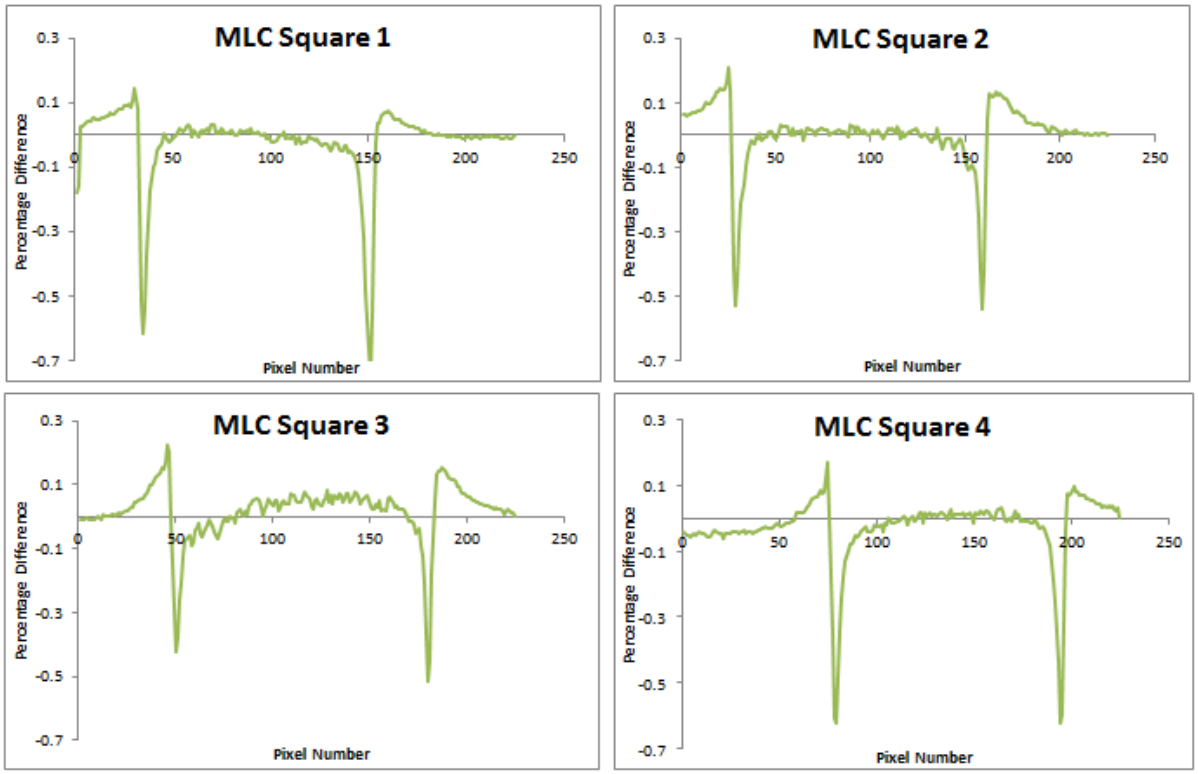

Figure 3.A.7 indicates that the TPS tends to slightly underestimate the film dose in the tail of the

penumbra, while it overestimates the dose in the shoulder of the penumbra. This matches what was seen in the time resolved sweeping window measurements (see **Figure 3.A.1**). The overestimation is more substantial in the film measurements as the beam penumbra is measured off axis, while the sweeping window measurements were all measured at isocentre. This also accounts for the higher deviation in MLC squares 1 and 4 (which are further off axis) compared to squares 2 and 3 (closer to central axis).

Similar plots were then created using the two different DLG values outlined earlier. These plots are included below (see **Figure 3.A.8**). It can be clearly seen that the agreement between film and TPS data is much better in the penumbra region for each MLC square using the 1.2 mm DLG compared to the 2.0 mm DLG (around ±0.1% for the 1.2 mm DLG compared to up to ±0.6% for the 2.0 mm DLG). Furthermore, the percentage differences are evenly distributed between the slight underestimate in the tail of the penumbra and the slight overestimate in the shoulder of the penumbra using the 1.2 mm DLG model. Therefore, any accumulated errors in the penumbra region will likely cancel out over the course of a VMAT plan where the MLC leaves are moving back and forth. Whereas for the 2.0 mm DLG model, the underestimate in the penumbra shoulder is larger than the overestimate in the tail, which may lead to a net underestimate of dose in the penumbra of the field. Therefore, the 1.2 mm DLG was selected as the optimal DLG setting.

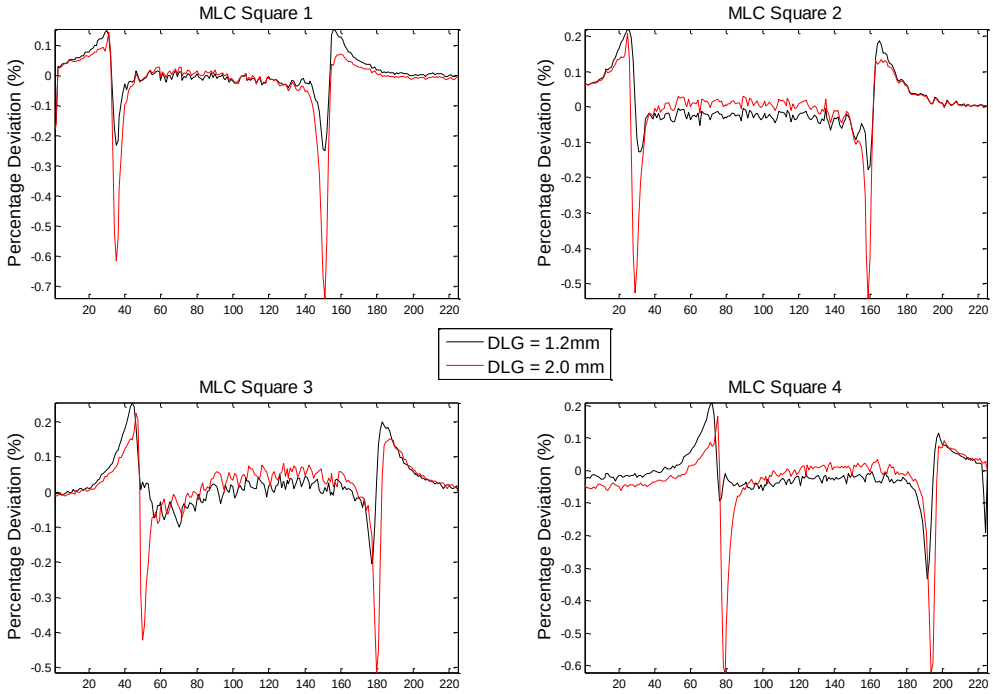

Figure 3.A.8: *Comparison profiles through each of the four MLC squares in the x-direction showing the difference between measured film dose and TPS calculated dose using the 1.2 mm DLG model (black) and the 2.0 mm DLG model (red). For each model the ETSS was 0.0 mm in both x and y directions. The y axis of each plot shows percentage difference while the x axis is pixel number.*

A similar comparison was then made for the three different ETSS settings for the x direction and these comparisons are plotted in **Figure 3.A.9**. As for the above sweeping window measurements, the ETSS of 0.0 mm in the x direction gave the worst agreement with measured data. Again, the difference between the deviations for the 1.0 mm and 1.5 mm ETSS were much more similar. The 1.0 mm ETSS resulted in a larger deviation in the tail of the penumbra but a smaller deviation in the shoulder compared to the 1.5 mm ETSS, although this was less obvious than for the sweeping window measurements. The deviations in the shoulder had the opposite sign compared to the deviations in the tail. If the deviations were of a similar magnitude they will tend to average out over many beam segments, resulting in a smaller net deviation. Therefore it was still concluded that the 1.5 mm ETSS gave the best agreement due to the more similar magnitudes of the deviations in the shoulder and tail of the penumbra.

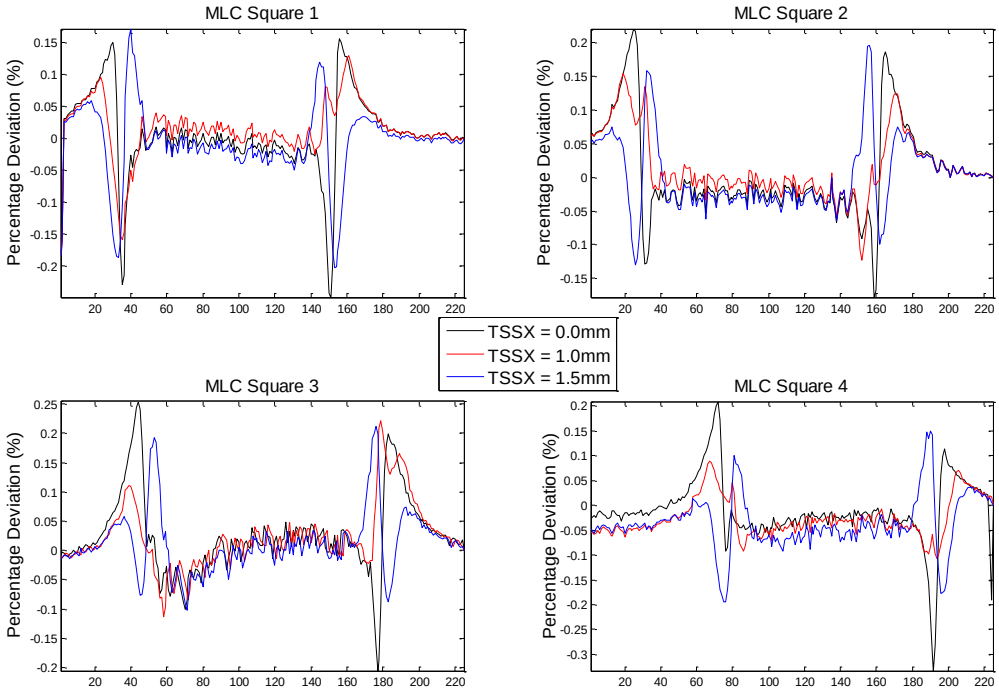

Figure 3.A.9: *Comparison profiles through each of the four MLC squares in the x-direction showing the difference between measured film dose and TPS calculated dose using the 0.0 mm ETSS X model (black), the 1.0 mm ETSS X model (red), and the 1.5 mm ETS SX model (blue). For each model the DLG was set to 1.2mm and the ETSS was 0.0 mm in the y direction. The y axis of each plot shows percentage difference while the x axis is pixel number.*

Therefore, it was concluded from the above results that the optimal beam model settings were a DLG value of 1.2 mm and an ETSS value of 1.5 mm in the x direction while retaining the 0.0 mm value in the y direction. The beam model using these values will henceforth be referred to as the **'adjusted beam model'**, while the beam model using the DLG value of 2.0 mm and ETSS values of 0.0 mm in both x and y directions will be referred to as the **'clinical beam model'**. All subsequent plans containing introduced errors were calculated using both beam models.

7. Bibliography

[1] World Health Organization, "WHO Cancer fact sheet," February 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed 12 May 2015].

[2] Ministry of Health, "Cancer: New registrations and deaths 2011," Ministry of Health, Wellington, 2014.

[3] R. Baskar, K. Lee, R. Yeo and K. Yeoh, "Cancer and Radiation Therapy: Current Advances and Future Directions," *Int. J. Med. Sci.*, vol. 9, no. 3, pp. 193 - 199, 2012.

[4] S. Webb and A. Nahum, "A model for calculating tumor control probability in radiotherapy includuing the effects of inhomogeneous distributions of dose and clonogenic cell density," *Phys. Med. Biol.*, vol. 38, no. 6, pp. 653 - 666, 1993.

[5] L. Marks, et al., "Use of normal tissue complication probability models in the clinic," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 76, no. 3, pp. S10 - S19, 2010.

[6] E. Hall and A. Giaccia, Radiobiology for the Radiobiologist, Philadelphia: Lippincott, Williams & Wilkins, 2010.

[7] Y. Yang and L. Xing, "Optimization of radiotherapy dose-time fractionation with consideration of tumor specific biology," *Med. Phys.*, vol. 32, no. 12, pp. 3666 - 3677, 2005.

[8] E. Podgorsak, Radiation Oncology Physics: A Handbook for Teachers and Students, Vienna: IAEA, 2005.

[9] I. El Naqa, P. Pater and J. Seuntjens, "Monte Carlo role in radiobiological modelling of radiotherapy outcomes," *Phys. Med. Biol.*, vol. 57, no. 11, pp. 75 - 97, 2012.

[10] International Atomic Energy Agency, "Transition from 2-D Radiotherapy to 3-D Conformal and Intensity Modulated Radiotherapy," International Atomic Energy Agency, Vienna, 2008.

[11] C. Ling and F. Fuks, "Conformal Radiation Treatment: a Critical Appraisal," *Eur. J. Cancer*, vol. 31A, no. 5, pp. 799 - 803, 1995.

[12] R. Mohan, "Field Shaping for Three-Dimensional Conformal Radiation Therapy and Multileaf Collimation," *Semin. Radiat. Oncol.*, vol. 5, no. 2, pp. 86 - 99, 1995.

[13] A. Oliveira, J. Vieira and F. Lima, "Monte Carlo Modeling of Multileaf Collimators Using the Code Geant4," *Braz. J. Rad. Sci.*, vol. 3, no. 1A, 2013.

[14] P. Xia and L. Verhey, "Multileaf Collimator leaf sequencing algorithm for intensity modulated beams with multiple static segments," *Med. Phys.*, vol. 25, no. 8, pp. 1624 - 1434, 1998.

[15] L. Verhey, "Comparison of Three-Dimensional Conformal Radiation Therapy and Intensity-Modulated Radiation Therapy Systems," *Semin. Radiat. Oncol.*, vol. 9, no. 1, pp. 78 - 98, 1999.

[16] T. Bortfeld, "IMRT: a review and preview," *Phys. Med. Biol.*, vol. 51, no. 13, pp. R363 - R379, 2006.

[17] F. Khan and R. Potish, Treatment Planning in Radiaiton Oncology, Maryland, USA: Williams & Wilkins, 1998.

[18] S. Webb, "The physical basis of IMRT and inverse planning," *Br. J. Radiol.,* vol. 76, pp. 678 - 689, 2003.

[19] C. Yu, "Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy," *Phys. Med. Biol.,* vol. 40, no. 9, pp. 1435 - 1449, 1995.

[20] K. Otto, "Volumetric modulated arc therapy: IMRT in a single gantry arc," *Med. Phys.,* vol. 35, no. 1, pp. 310 - 317, 2008.

[21] A. Holt, et al., "Multi-institutional comparison of volumetric modulated arc therapy vs. intensity-modulated radiation therapy for head-and-neck cancer: a planning study," *Radiat. Oncol.,* vol. 8, no. 26, 2013.

[22] International Commission on Radiation Units and Measurements, "Report No. 62: Prescribing, Recording and Reporting Photon Beam Therapy," International Commission on Radiation Units and Measurements, Bethesda, 1999.

[23] B. Mijnheer, J. Battermann and A. Wambersie, "What degree of accuracy is required and can be achieved in photon and neutron therapy?," *Radiother. Oncol.,* vol. 8, pp. 237 - 252, 1987.

[24] D. Thwaites, "Accuracy required and achievable in radiotherapy dosimetry: have modern technology and techniques changed our views?," *J. Phys. Conf. Ser.,* vol. 444, p. 012006, 2013.

[25] G. Hartmann, "Quality Management in Radiotherapy," in *New Technologies in Radiation Oncology*, Berlin, Springer-Verlag, 2006, pp. 425 - 447.

[26] W. Mayles, et al., "Physics Aspects of Quality Control in Radiotherapy," The Institute of Physics and Engineering in Medicine, York, 1999.

[27] G. Kutcher, et al., "Comprehensive QA for radiation oncology: Report of AAPM Radiation Therapy Committee Task Group 40," American Association of Physicists in Medicine, College Park, 1994.

[28] M. Alber, et al., "ESTRO Booklet No. 9 Guidelines on the Verification of IMRT," European Society for Theraputic Radiology and Oncology, Brussels, 2008.

[29] Netherlands Commission on Radiation Dosimetry, "Report 24: Code of Practise for the Quality Assurance and Control for Volumetric Modulated Arc Therapy," Netherlands Commission on Radiation Dosimetry, Delft, 2015.

[30] Netherland Commission on Radiation Dosimetry, "Code of Practise for the Quality Assurance and Control for Intensity Modulated Radiotherapy," Netherland Commission on Radiation Dosimetry, Delft, 2013.

[31] R. Macklis, T. Meier and M. Weinhous, "Error rates in clinical radiotherapy," *J. Clin. Oncol.,* vol. 16, no. 2, pp. 551 - 556, 1998.

[32] D. Margalit, et al., "Technological Advancements and Error Rates in Radiation Therapy Delivery," *Int. J. Radiat. Oncol. Biol. Phys.,* vol. 81, no. 4, pp. 673 - 679, 2011.

[33] United States Nuclear Regulatory Commission, “Gamma knife treatment to the wrong side of brain. Event Notification Report 43746,” Detroit, 2007.

[34] Scottish Ministers for the Ionising Radiation (Medical Exposures) Regulations 2000, “Unintended overexposure of patient Lisa Norris during radiotherapy treatment at the Beatson Oncology Centre, Glasgow in January 2006,” Edinburgh, 2006.

[35] International Atomic Energy Agency, “Investigation of an accidental exposure of radiotherapy patients in Panama: Report of a team of experts,” International Atomic Energy Agency, Vienna, 2001.

[36] International Atomic Energy Agency, “Accidental overexposure of radiotherapy patients in San Jose, Costa Rica,” International Atomic Energy Agency, Vienna, 1998.

[37] International Atomic Energy Agency, “Accidental Overexposure of radiotherapy patients in Bialystok,” International Atomic Energy Agency, Vienna, 2004.

[38] International Commission of Radiation Protection, “ICRP 112: Preventing Accidental Exposures from New External Beam Radiation Therapy Technologies,” International Commission of Radiation Protection, 2009.

[39] Interntaional Atomic Energy Agency, “SRS-17: Lessons learned from accidental exposures in radiotherapy,” Interntaional Atomic Energy Agency, Vienna, 2000.

[40] B. Fraas, “Errors in Radiotherapy: Motivation for development of new radiotherapy quality assurance paradigms,” *Int. J. Radiat. Oncol. Biol. Phys.,* vol. 71, no. 1, pp. S162 - S165, 2008.

[41] T. Yeung, et al., “Quality Assurance in Radiotherapy: evaluation of errors and incidents recorded over a 10 year period,” *Radiother. Oncol.,* vol. 74, pp. 283 - 291, 2004.

[42] International Organization for Standardization/International Electrotechnical Commission, “International vocabulary for metrology - Basic and general concepts and associated terms (VIM): Guide 99,” International Organization for Standardization/International Electrotechnical Commission, Switzerland, 2007.

[43] B. Nelms, et al., “MO-EE-A2_01: On the Dosimetric/DVH Impact of Variation in Organ Delineation: A Multi-Institutional Study and Proposed Quality System,” *Med. Phys.,* vol. 37, no. 6, pp. 3348 - 3348, 2010.

[44] J. Kim, et al., “The sensitivity of gamma-index method to the positioning errors of high-definnition MLC in patient-specific VMAT QA for SBRT,” *Radiat. Oncol.,* vol. 9, no. 1, pp. 167 - 179, 2014.

[45] C. Bojechko and E. Ford, “Quantifying the performance of in vivo portal dosimetry in detecting four types of treatment parameter variations,” *Med. Phys.,* vol. 42, no. 12, pp. 6912 - 6918, 2015.

[46] M. Carlone, et al., “ROC analysis in patient specific quality assurance,” *Med. Phys.,* vol. 40, no. 4, pp. 42103-1 - 42103-7, 2013.

[47] S. Kry, et al., “Patient-Specific IMRT QA does not predict unacceptable plan delivery,” *Int. J. Radiat. Oncol. Biol. Phys.,* vol. 90, pp. 1195-1201, 2014.

[48] M. Oliver, et al., “Clinical significance of multi-leaf collimator positional errors for volumetric

modulated arc therapy," *Radiot. Oncol.,* vol. 97, pp. 554 - 560, 2010.

[49] A. Fredh, et al., "Patient QA systems for rotational radiation therapy: A comparative study with intentional errors," *Med. Phys.,* vol. 40, no. 3, p. 031716, 2013.

[50] B. Nelms, H. Zhen and W. Tome, "Per-beam, planar IMRT QA passing rates do not predict clincially relevant patient dose errors," *Med. Phys.,* vol. 38, no. 2, pp. 1037-1044, 2011.

[51] R. Louwe, et al., "Time resolved dosimetry using a pinpoint ionization chamber as quality assurance for IMRT and VMAT," *Med. Phys.,* vol. 42, no. 4, pp. 1625 - 1639, 2015.

[52] J. Kruse, "On the insensitivity of single field planar dosimetry to IMRT inaccuracies," *Med. Phys.,* vol. 37, no. 6, pp. 2516 - 2524, 2010.

[53] V. Gregoire, K. Ang, W. Budach, et al., "Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines," *Radiot. Oncol.,* vol. 110, no. 1, pp. 172 - 181, 2014.

[54] International Commissioin on Radiation Units and Measurements, "ICRU Report 83: Prescribing, Recordng, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT)," *J. ICRU,* vol. 10, no. 1, 2010.

[55] Varian Medical Systems, "Eclispe Algorithms Reference Guide," Varian Medical Systems, Palo Alto, 2011.

[56] WBCC, "Planning - RapidArc for head and neck," WBCC internal report, Wellington, 2015.

[57] M. Huq, et al., "The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management," *Med. Phys.,* vol. 43, no. 7, pp. 4209 - 4262, 2016.

[58] E. Klein, et al., "Errors in radiaiton oncology: A study in pathways and dosimetric impact," *J. Appl. Clin. Med. Phys.,* vol. 6, no. 3, pp. 81 - 94, 2005.

[59] J. Kerns, N. Childress and S. Kry, "A multi-institution evaluation of MLC log files and performance in IMRT delivery," *Radiat. Oncol.,* vol. 176, no. 9, 2014.

[60] S. Goetsch, *Patient Safety and Radiation Accidents,* RTUVT University, South Bend, 2013.

[61] G. Heilemann, B. Poppe and W. Laub, "On the sensitivity of common gamma-index evaluation methods to MLC misalignments," *Med. Phys.,* vol. 40, no. 3, pp. 031702-1 - 12, 2013.

[62] T. Kron, et al., "Small field segments surrounded by large areas only shielded by a multileaf collimator: Comparison of experiments and dose calculation," *Med. Phys.,* vol. 39, no. 12, pp. 7480 - 7489, 2012.

[63] A Fogliata, et al., "Accuracy of Acuros XB and AAA dose calculation for small fields with reference to RapidArc stereotactic treatments," *Med. Phys.,* vol. 38, no. 11, pp. 6228 - 6237, 2011.

[64] B. Steer, "Difference in DLG between LA1 and LA2," WBCC Internal Report, Wellington, 2015.

[65] J. Chow, G. Grigorov and R. Jiang, “Intensity modulated radiation therapy with irregular multileaf collimated field: A dosimetric study on the penumbra region with different leaf stepping patterns,” *Med. Phys.,* vol. 33, no. 12, pp. 4606 - 4613, 2006.

[66] PTW, “User Manual PinPoint Chambers,” PTW-Freiburg, Freiburg, 2008.

[67] International Atomic Energy Agency, “TRS-398 Absorbed Dose Determination in External Beam Radiotherapy,” International Atomic Energy Agency, Vienna, 2006.

[68] T. Satherley, “Tandem Software User Guide,” WBCC Internal Document, Wellington, 2014.

[69] C. Andres, et al., “A comprehensive study of the Gafchromic EBT2 radiochromic film. A comparison with EBT,” *Med. Phys.,* vol. 37, no. 12, pp. 6271 - 6278, 2010.

[70] B. Ferreira, M. Lopes and M. Caoela, “Evaluation of an Epson flatbed scanner to read Gafchromic EBT films for radiation dosimetry,” *Phys. Med. Biol.,* vol. 54, pp. 1073 - 1085, 2009.

[71] F. Del Moral, et al., “From the limits of the classical model of sensitometric curves to a realistic model based on the percolation theory for GafChromic EBT films,” *Med. Phys.,* vol. 36, no. 9, p. 4015, 2009.

[72] K. Levenberg, “A method for the solution of certain problems in least squares,” *Quart. Appl. Math.,* vol. 2, pp. 164-168, 1944.

[73] D. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *J. Soc. Indust. Appl. Math.,* vol. 11, no. 2, pp. 431 - 441, 1963.

[74] D. Low, “Gamma dose distribution evaluation tool,” *J. Phys. Conf. Ser.,* vol. 250, no. 1, p. 012071, 2010.

[75] D. Letourneac, J. Publicover, J. Kozelka, D. Moseley and D. Jaffray, “Novel dosimetric phantom for quality assurance of volumetric modulated arc therapy,” *Med. Phys.,* vol. 36, no. 5, pp. 1813 - 1821, 2009.

[76] Sun Nuclear Corporation, “ArcCheck Reference Guide,” Sun Nuclear Corperation, Melbourne, FL, 2015.

[77] L. Moran, “AC measurements with HUo and HC,” WBCC internal Report, Wellington, 2014.

[78] C. Mathias, “Sensitivity,” in *Encyclopedia of Research Design*, Thousand Oaks, CA, SAGE Publications Inc., 2010, pp. 1338-1340.

[79] M. Greiner, D. Pfeiffer and R. Smith, “Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests,” *Prev. Vet. Med.,* vol. 45, pp. 23 - 41, 2000.

[80] J. Hanley and B. McNeil, “The meaning and Use of the Area under a Reciever Operating Characterisitic (ROC) Curve,” *Radiology,* vol. 143, pp. 29 - 36, 1982.

[81] K. Hajian-Tilaki, “Reciever Operating Characterisitic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” *Caspian. J. Intern. Med.,* vol. 4, no. 2, pp. 627 - 635, 2013.

[82] N. Perkins and E. Schisterman, “The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve,” *Am. J. Epidemiol.,* vol. 163, no.

7, pp. 670 - 675, 2005.

[83] G. Yan, et al., "On the sensitivity patient-specific IMRT QA to MLC positioning errors," *J. App. Clin. Med. Phys.,* vol. 10, no. 1, pp. 120 - 128, 2009.

[84] R. Louwe, T. Satherley, A. Williams and B. Scarlet, *Optimisation of TPS beam model parameters for stereotactic treatments using VMAT,* Christchurch: Oral Presentation at the NZ SABR workshop, 21 May 2016.

[85] E. McKenzie, et al., "Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans," *Med. Phys.,* vol. 41, no. 12, p. 121702, 2014.

[86] M. Aristophanous, et al., "Initial clinical experience with ArcCHECK for IMRT/VMAT QA," *J. Clin. App. Med. Phys.,* vol. 17, no. 5, pp. 20 - 33, 2016.

[87] L. Coleman and C. Skourou, "Sensitivity of volumetric modulated arc therapy patient specific QA results to multileaf collimator errors and correlation to dose volume histogram based metrics," *Med. Phys.,* vol. 40, no. 11, pp. 111715 - 111722, 2013.

[88] S. Vida, "A computer program for non-parametric receiver operating characteristic analysis," *Comput. Methods Programs Biomed.,* vol. 40, pp. 95 - 101, 1993.

[89] G. Ezzell, et al., "IMRT commissioing: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," AAPM, Phoenix, USA, 2009.

[90] B. Nelms, et al., "Evaluating IMRT and VMAT dose accuracy: Practical examples of failure to detect systematic errors when applying a commonly used metric and action levels," *Med. Phys.,* vol. 40, no. 11, 2013.

[91] K. Kielar, E. Mok, A. Hsu, L. Wang and G. Luxtion, "Verification of dosimetric accuracy of the TrueBeam STx: Rounded leaf effect of the high definition MLC," *Med. Phys.,* vol. 39, no. 10, pp. 6360 - 6371, 2012.

[92] T. LoSasso, C. Chui and C. Ling, "Physical and dosimetric aspects of a multileaf collimation system used in the dynamic mode for implementing intensity modulated radiotherapy," *Med. Phys.,* vol. 25, no. 10, pp. 1919 - 1927, 1998.

[93] X. Mei, I. Nygren and J. Villarreal-Barajas, "On the use of the MLC dosimetric leaf gap as a quality control tool for accurate dynamic IMRT delivery," *Med. Phys.,* vol. 38, no. 4, pp. 2246 - 2255, 2011.

[94] A. Williams, R. Louwe and T. Satherley, *Optimisation of Eclipse Beam Model of VMAT Delivery,* Wellington: Oral Presentation at EPSM2015, 2015.